1-2013

Variable Selection in Nonparametric and Semiparametric Regression Models

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

Yonghui ZHANG

Variable Selection in Nonparametric and Semiparametric Regression Models*

Liangjun Su, Yonghui Zhang
School of Economics, Singapore Management University

September 19, 2012

Abstract

This chapter reviews the literature on variable selection in nonparametric and semiparametric regression models via shrinkage. We highlight recent developments on simultaneous variable selection and estimation through the methods of *least absolute shrinkage and selection operator* (Lasso), *smoothly clipped absolute deviation* (SCAD) or their variants, but restrict our attention to nonparametric and semiparametric regression models. In particular, we consider variable selection in additive models, partially linear models, functional/varying coefficient models, single index models, general nonparametric regression models, and semiparametric/nonparametric quantile regression models.

JEL Classifications: C14, C52

Key Words: Cross validation, High dimensionality, Lasso, Nonparametric regression, Oracle property, Penalized least squares, Penalized likelihood, SCAD, Semiparametric regression, Sparsity, Variable selection

1 Introduction

Over the last 15 years or so high dimensional models have become increasingly popular and gained considerable attention in diverse fields of scientific research. Examples in economics include wage regression with more than 100 regressors (e.g., Belloni and Chernozhukov, 2011b), portfolio allocation among hundreds or thousands of stocks (e.g., Jagannathan and Ma, 2003; Fan, Zhang and Yu, 2011), VAR models with hundreds of data series (e.g., Bernanke, Boivin and Eliasz, 2005), large dimensional panel data models of home price (e.g., Fan, Lv and Qi, 2011), among others. A common feature of high dimensional models is the number of regressors is very large, which may grow as the sample size increases. This poses a series of challenges for statistical modeling and inference. Penalized

*Address Correspondence to: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; E-mail: ljsu@smu.edu.sg, Phone: +65 6828 0386.

least squares or likelihood has become a standard unified framework for variable selection and feature extraction in such scenarios. For a comprehensive overview of high dimensional modeling, see Fan and Li (2006) and Fan and Lv (2010).

In high dimensional modeling, one of the most important problems is the choice of an optimal model from a set of candidate models. In many cases, this reduces to the choice of which subset of variables should be included into the model. Subset selection has attracted a lot of interest, leading to a variety of procedures. The majority of these procedures do variable selection by minimizing a penalized objective function with the following form:

$$\text{Loss function} \ + \ \text{Penalization}.$$

The most popular choices of loss functions are least squares, negative log-likelihood, and their variants (e.g., profiled least squares or profiled negative log-likelihood). There are many choices of penalization. The traditional subset selection criterion such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) uses the l_0-norm for the parameters entering the model so that the penalization term is proportional to the number of nonzero parameters. The bridge estimator (see, e.g., Frank and Friedman, 1993; Fu, 1998; Knight and Fu, 2000) uses the l_q-norm ($q > 0$). It boils down the commonly used ridge estimator (Hoerl and Kennard, 1970) when $q = 2$ and the Lasso estimator (Tibshirani, 1996) when $q = 1$. When $0 < q \leq 1$, some components of the estimator can be exactly zero with some correctly chosen tuning parameters. Thus, the bridge estimator with $0 < q \leq 1$ provides a way to combine variable selection and parameter estimation simultaneously. Among the class of bridge estimators, Lasso becomes most popular due to its computational and theoretical advantages comparing with other bridge estimators and traditional variable selection methods. Allowing an adaptive amount of shrinkage for each regression coefficient results in an estimator called the adaptive Lasso that is first proposed by (Zou, 2006) and can be as efficient as the oracle one in the sense that the method works asymptotically equivalent to the case as if the correct model were exactly known. Other variants of Lasso include the group Lasso, adaptive group Lasso, graphic Lasso, elastic net, etc. Of course, the penalization term can take other forms; examples include the SCAD penalty of Fan and Li (2001) and the MC penalty of Zhang (2010).

Given the huge literature on variable selection that has developed over the last 15 years, it is impossible to review all of the works. Fan and Lv (2010) offer a selective overview of variable selection in high dimensional feature space. By contrast, in this chapter we focus on variable selection in semiparametric and nonparametric regression models with fixed or large dimensions because semiparametric and nonparametric regressions have gained considerable importance over the last three decades due to their flexibility in modeling and robustness to model misspecification. In particular, we consider variable selection in the following models

- additive models
- partially linear models
- functional/varying coefficient models

- single index models
- general nonparametric regression models
- semiparametric/nonparametric quantile regression models

The first four areas are limited to semiparametric and nonparametric regression models that impose certain structure to alleviate the "curse of dimensionality" in the nonparametric literature. The fifth part focuses on variable or component selection in general nonparametric models. In the last part we review variable selection in semiparametric and nonparametric quantile regression models. Below we first briefly introduce variable selection via Lasso or SCAD type of penalties in general parametric regression models and then review its development in the above fields in turn. In the last section we highlight some issues that require further study.

2 Variable Selection via Lasso or SCAD Type of Penalties in Parametric Models

In this section we introduce the background of variable selection via Lasso or SCAD type of penalties.

2.1 The Lasso estimator

The Lasso (*least absolute shrinkage and selection operator*) proposed by Tibshirani (1996) is a popular model building technique which can simultaneously produce accurate and parsimonious models. For $i = 1, ..., n$, let Y_i denote a response variable and $X_i = (X_{i1}, ..., X_{ip})'$ a $p \times 1$ vector of covariates/predictors. For simplicity, one could assume that (X_i, Y_i) are independent and identically distributed (i.i.d.), or assume that $\{X_i, Y_i\}_{i=1}^n$ are standardized so that $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i = 0$, $n^{-1}\sum_{i=1}^n X_{ij} = 0$, and $n^{-1}\sum_i X_{ij}^2 = 1$ for $j = 1, ..., p$. But these are not necessary. The Lasso estimates of the slope coefficients in a linear regression model solve the l_1-penalized least regression problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \beta_j X_{ij} \right)^2 \text{ subject to } \sum_{j=1}^p \left|\beta_j\right| \leq s, \tag{2.1}$$

or equivalently,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p \left|\beta_j\right|, \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, and s and λ are tuning parameters. (2.1) suggests that the Lasso uses a constraint in the form of l_1-norm: $\sum_{j=1}^p \left|\beta_j\right| \leq s$. It is similar to the ridge regression with the constraint of l_2-norm: $\sum_{j=1}^p \beta_j^2 \leq s$. By using the l_1-penalty, the Lasso achieves variable selection and shrinkage simultaneously. However, ridge regression only does shrinkage. More generally, a *penalized least squares* (PLS) can have a generic l_q-penalty of the form $(\sum_{j=1}^p \left|\beta_j\right|^q)^{1/q}$, $0 \leq q \leq 2$. When $q \leq 1$, the PLS automatically performs variable selection by removing predictors with very small estimated coefficients.

In particular, when $q = 0$, the l_0-penalty term becomes $\sum_{j=1}^{p} \mathbf{1}\left(\beta_j \neq 0\right)$ with $\mathbf{1}(\cdot)$ being the usual indicator function, which counts the number of nonzero elements in the vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$. The Lasso uses the smallest value of q, i.e., $q = 1$, which leads to a convex problem.

The tuning parameter λ is the shrinkage parameter that controls the amount of regularization. If $\lambda = 0$, there is no penalty put on the coefficients and hence we obtain the ordinary least squares solution; if $\lambda \to \infty$, the penalty is infinitely large and thus forces all of the coefficients to be zero. These are necessary but insufficient for the Lasso to produce sparse solutions. Large enough λ will set some coefficients exactly equal to zero and is thus able to perform variable selection. In contrast, a ridge penalty never sets coefficients exactly equal to 0.

Efron, Hastie, Johnstone and Tibshirani (2004) propose the *least angel regression selection* (LARS) and show that the entire solution path of the Lasso can be computed by the LARS algorithm. Their procedure includes two steps. First, a solution path which is indexed by a tuning parameter is constructed. Then the final model is chosen on the solution path by cross-validation or using some criterion such as C_p. The solution path is piecewise linear and can be computed very efficiently. These nice properties make the Lasso very popular in variable selection.

2.2 Some generalizations and variants of the Lasso

In this subsection, we review some variants of the Lasso: Bridge, the adaptive Lasso, the group Lasso, and the elastic-net. For other work generalizing the Lasso, we give a partial list for reference.

Bridge. Knight and Fu (2000) study the asymptotics for the Bridge estimator $\hat{\boldsymbol{\beta}}_{\text{Bridge}}$ which is obtained via the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} \left| \beta_j \right|^{\gamma}, \tag{2.3}$$

where λ_n is a tuning parameter and $\gamma > 0$. For $\gamma \leq 1$, the Bridge estimator has the potentially attractive feature of being exactly 0 if λ_n is sufficiently large, thus combining parametric estimation and variable selection in a single step.

Adaptive Lasso. Fan and Li (2001) and Fan and Peng (2004) argue that a good variable selection procedure should have the following oracle properties: (i) (selection consistency) it can identify the right subset models in the sense that it selects the correct subset of predictors with probability tending to one, and (ii) (asymptotic optimality) it achieves the optimal asymptotic distribution as the oracle estimator in the sense that it estimates the nonzero parameters as efficiently as would be possible if we knew which variables were uninformative ahead of time. It has been shown that the Lasso of Tibshirani (1996) lacks such oracle properties whereas the Bridge estimator with $0 < \gamma < 1$ can possess them with a well-chosen tuning parameter. Fan and Li (2001) point out that asymptotically the Lasso has non-ignorable bias for estimating the nonzero coefficients. Zou (2006) first shows that the Lasso could be inconsistent for model selection unless the predictor matrix satisfies a rather strong condition, and then propose a new version of the Lasso, called the *adaptive Lasso*. Adaptive Lasso assigns different weights to penalize different coefficients in the l_1-penalty. That is, the adaptive Lasso estimate $\hat{\boldsymbol{\beta}}_{\text{ALasso}}$

of $\boldsymbol{\beta}$ solves the following minimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left| \beta_j \right|, \tag{2.4}$$

where the weights $\hat{w}_j$'s are data-dependent and typically chosen as $\hat{w}_j = |\hat{\beta}_j|^{-\gamma}$ for some $\gamma > 0$, and $\hat{\beta}_j$ is a preliminary consistent estimator of β_j that typically has $\sqrt{n}$-rate of convergence. Intuitively, in adaptive Lasso the zero parameter is penalized more severely than a non-zero parameter as the weight of the zero parameter goes to infinity while that of a nonzero parameter goes to a positive constant. Zou shows that the adaptive Lasso enjoys the oracle properties so that it performs as well as if the underlying true model were given in advance. Similar to the Lasso, the adaptive Lasso is also near-minimax optimal in the sense of Donoho and Johnstone (1994).

Group Lasso. Observing that the Lasso is designed for selecting individual regressor, Yuan and Lin (2006) consider extensions of the Lasso and the LARS to the case with "grouped variables". The group Lasso estimate is defined as the solution to

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \mathbf{Y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda_n \sum_{j=1}^{J} \left\| \boldsymbol{\beta}_j \right\|_{K_j}, \tag{2.5}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)'$, $\mathbf{X}_j$ is an $n \times p_j$ regressor matrix, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_J')'$, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector, K_j is a $p_j \times p_j$ positive definite matrix, and $\|\cdot\|_2$ is an Euclidean norm, $\left\| \boldsymbol{\beta}_j \right\|_{K_j} = (\boldsymbol{\beta}_j' K_j \boldsymbol{\beta}_j)^{1/2}$, and $\lambda \geq 0$ is a tuning parameter. Two typical choices of K_j are I_{p_j} and $p_j I_{p_j}$, where I_{p_j} is a $p_j \times p_j$ identity matrix and the coefficient p_j in the second choice is used to adjust for the group size. Obviously, the penalty function in the group Lasso is intermediate between the l_1-penalty that is used in the Lasso and the l_2-penalty that is used in ridge regression. It can be viewed as an l_1-penalty used for coefficients from different groups and an l_2-penalty used for coefficients in the same group. Yuan and Lin propose a group version of the LARS algorithm to solve the minimization problem. See Huang, Breheny, and Ma (2012) for an overview on group selection in high dimensional models.

Elastic-net. As shown in Zou and Hastie (2005), the Lasso solution paths are unstable when the predictors are highly correlated. Zou and Hastie (2005) propose the elastic-net as an improved version of the Lasso for analyzing high-dimensional data. The elastic-net estimator is defined as follows

$$\hat{\boldsymbol{\beta}}_{\text{Enet}} = \left(1 + \frac{\lambda_2}{n} \right) \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\}, \tag{2.6}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)'$, $\mathbf{X}$ is an $n \times p$ regressor matrix, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, $\|\boldsymbol{\beta}\|_q = \{\sum_{j=1}^{p} \left|\beta_j\right|^q\}^{1/q}$, and λ_1 and λ_2 are tuning parameters. When the predictors are standardized, $1+\lambda_2/n$ should be replaced by $1+\lambda_2$. The l_1-part of the elastic-net performs automatic variable selection, while the l_2-part stabilizes the solution paths to improve the prediction. Donoho, Johnstone, Kerkyacharian, and Picard (1995) show that in the case of orthogonal design the elastic-net automatically reduces to the Lasso. Zou and Hastie also propose the adaptive elastic-net estimates:

$$\hat{\boldsymbol{\beta}}_{\text{AdaEnet}} = \left(1 + \frac{\lambda_2}{n} \right) \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} \hat{w}_j \left| \beta_j \right| \right\} \tag{2.7}$$

where $\hat{w}_j = |\hat{\beta}_{\text{Enet},j}|^{-\gamma}$ and $\hat{\beta}_{\text{Enet},j}$ denotes the jth element of $\hat{\boldsymbol{\beta}}_{\text{Enet}}$ for $j = 1, ..., p$. Under some weak regularity conditions, they establish the oracle property of the adaptive elastic-net.

There has been a large amount of work in recent years, applying and generalizing the Lasso and l_1-like penalties to a variety of problems. This includes the adaptive group Lasso (e.g., Wang and Leng, 2008; Wei and Huang, 2010), fused Lasso (e.g., Tibshirani et al., 2005; Rinaldao, 2009), the graphical Lasso (e.g., Yuan and Lin, 2007; Friedman et al., 2008), the Dantzig selector (e.g., Candès and Tao, 2007), near isotonic regularization (e.g., Tibshirani et al., 2010), among others. See Table 1 in Tibshirani (2011) for a partial list of generalizations of the Lasso.

2.3 Other penalty functions

Many non-Lasso-type penalization approaches have also been proposed, including the SCAD and MC penalties. In the linear regression framework, the estimates are given by

$$\hat{\boldsymbol{\beta}}_n(\lambda_n) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{p} p_{\lambda_n}\left(|\beta_j|\right),$$

where $p_{\lambda_n}(\cdot)$ is a penalty function and λ_n is a penalty parameter. Different penalty functions yield different variable selection procedures, which have different asymptomatic properties. Note that the Bridge penalty function takes the form $p_\lambda(v) = \lambda|v|^\gamma$, where $\gamma > 0$ is a constant. The ordinary Lasso penalty function corresponds to $p_\lambda(v) = \lambda|v|$.

SCAD. The SCAD (*smoothly clipped absolute deviation*) penalty function proposed by Fan and Li (2001) takes the form

$$p_\lambda(v) = \begin{cases} \lambda v & \text{if } 0 \le v \le \lambda \\ -\frac{\left(v^2 - 2a\lambda v + \lambda^2\right)}{2(a-1)} & \text{if } \lambda < v < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } a\lambda \le v, \end{cases}$$

and its derivative satisfies that

$$p'_\lambda(v) = \lambda\left[1(v \le \lambda) + \frac{(a\lambda - v)_+}{(a-1)\lambda}1(v > \lambda)\right]$$

where $(b)_+ = \max(b, 0)$ and $a > 2$ is a constant ($a = 3.7$ is recommended and used frequently). Fan and Li (2001) and Fan and Peng (2004) investigate the properties of penalized least squares and likelihood estimator with the SCAD penalty. In particular, they show that the SCAD penalty can yield estimators with the oracle properties. Hunter and Li (2004) suggest using MM algorithms to improve the performance of SCAD-penalized estimators.

MC. The MC (*minimax concave*) penalty function proposed by Zhang (2010) is given

$$p_\lambda(v) = \lambda\int_0^v \left(1 - \frac{x}{\gamma\lambda}\right)_+ dx$$

where $\gamma > 0$ is a tuning parameter. Zhang proposes and studies the MC+ methodology that has two components: a MC penalty and a penalized linear unbiased selection (PLUS) algorithm. It provides

a fast algorithm for nearly unbiased concave penalized selection in the linear regression model and achieves selection consistency and minimax convergence rates.

For other penalization methods, see Fan, Huang and Peng (2005) and Zou and Zhang (2009).

3 Variable Selection in Additive Models

In this section, we consider the problem of variable selection in the following nonparametric additive model

$$Y_i = \mu + \sum_{j=1}^{p} f_j\left(X_{ij}\right) + \varepsilon_i, \tag{3.1}$$

where Y_i is a response variable and $X_i = (X_{i1}, ..., X_{ip})'$ is a $p \times 1$ vector of covariates, μ is the intercept term, the f_j's are unknown smooth functions with zero means, and ε_i is the unobserved random error term with mean zero and finite variance σ^2. It is typically assumed that $\left(Y_i, X_i\right), i = 1, ..., n$, are i.i.d. and some additive components $f_j\left(\cdot\right)$ are zero. The main problem is to distinguish the nonzero components from the zero components and estimate the nonzero components consistently. The number of additive components p can be either fixed or divergent as the sample size n increases. In the latter case, we frequently write $p = p_n$. In some scenarios, p_n is allowed to be much larger than n.

Many penalized methods have been proposed to select the significant nonzero components for model (3.1). Huang, Horowitz, and Wei (2010) apply the *adaptive group Lasso* to select nonzero component after using the group Lasso to obtain an initial estimator and to achieve an initial reduction of the dimension. They assume that the number of nonzero f_j's is fixed and give conditions under which the group Lasso selects a model whose number of components is comparable with the true model. They show that the adaptive group Lasso can select the nonzero components correctly with probability approaching one as n increases and achieves the optimal rates of convergence in the "large p, small n" setting. Meier, van de Geer and Bühlmann (2009) consider an additive model which allows for both the numbers of zero and nonzero f_j's passing to infinity and being larger than n. They propose a sparsity-smoothness penalty for model selection and estimation. Under some conditions, they show that the nonzero components can be selected with probability approaching 1. However the model selection consistency is not established. Ravikumar et al. (2009) propose a penalized approach for variable selection in nonparametric additive models. They impose penalty on the l_2-norm of the nonparametric components. Under some strong conditions on the design matrix and with special basis functions, they establish the model selection consistency. In all the above three approaches, the penalties are in the form of group/adaptive Lasso, or variants of the group Lasso. In addition, Xue (2009) proposes a penalized polynomial spline method for simultaneous variable selection and model estimation in additive models by using the SCAD penalty. We will review these methods in turn.

Several other papers have also considered variable selection in nonparametric additive models. Bach (2008) applies a method similar to the group Lasso to select variables in additive models with a fixed number of covariates and establishes the model selection consistency under a set of complicated conditions. Avalos, Grandvalet, and Ambroise (2007) propose a method for function estimation and

variable selection for additive models fitted by cubic splines, but they don't give any theoretic analysis for their method. Lin and Zhang (2006) propose the component selection and smoothing operator (COSSO) method for model selection and estimation in multivariate nonparametric regression models, in the framework of smoothing spline ANOVA. The COSSO is a method of regularization with the penalty functional being the sum of component norms, instead of the squared norm employed in the traditional smoothing spline method. They show that in the special case of a tensor product design, the COSSO correctly selects the nonzero additive component with high probability. More recently, Fan, Feng, and Song (2011) propose several closely related variable screening procedures in sparse ultral-high-dimensional additive models.

3.1 Huang, Horowitz, and Wei's (2010) adaptive group Lasso

Huang, Horowitz, and Wei (2010) propose a two-step approach to select and estimate the nonzero components simultaneously in (3.1) when p is fixed. It uses the group Lasso in the first stage and the adaptive group Lasso in the second stage.

Suppose that X_{ij} takes values in $[a,b]$ where $a<b$ are finite numbers. Suppose $E\left[f_j\left(X_{ij}\right)\right]=0$ for $j=1,...,p$ to ensure unique identification of f_j's. Under some suitable smoothness assumptions, f_j's can be well approximated by functions in $\mathcal{S}_n$, a space of polynomial splines defined on $[a,b]$ with some restrictions. There exists a normalized B-spline basis $\{\phi_k, 1\leq k\leq m_n\}$ for $\mathcal{S}_n$ such that for any $f_{nj}\in\mathcal{S}_n$, we have

$$f_{nj}\left(x\right)=\sum_{k=1}^{m_n}\beta_{jk}\phi_k\left(x\right),\ 1\leq j\leq p. \tag{3.2}$$

Let $\boldsymbol{\beta}_{nj}=\left(\beta_{j1},...,\beta_{jm_n}\right)'$ and $\boldsymbol{\beta}_n=(\boldsymbol{\beta}_{n1}',...,\boldsymbol{\beta}_{np}')'$. Let $w_n=(w_{n1},...,w_{np})'$ be a weight vector and $0\leq w_{nj}\leq\infty$ for $j=1,...,p$. Huang, Horowitz, and Wei consider the following penalized least squares (PLS) criterion

$$L_n\left(\mu,\boldsymbol{\beta}_n\right)=\sum_{i=1}^{n}\left[Y_i-\mu-\sum_{j=1}^{p}\sum_{k=1}^{m_n}\beta_{jk}\phi_k\left(X_{ij}\right)\right]^2+\lambda_n\sum_{j=1}^{p}w_{nj}\left\|\boldsymbol{\beta}_{nj}\right\|_2 \tag{3.3}$$

subject to

$$\sum_{i=1}^{n}\sum_{k=1}^{m_n}\beta_{jk}\phi_k\left(X_{ij}\right)=0,\ j=1,...,p, \tag{3.4}$$

where λ_n is a tuning parameter.

Note that (3.3) and (3.4) define a constrained minimization problem. To convert it to an unconstrained optimization problem, one can center the response and the basis functions. Let $\bar{\phi}_{jk}=n^{-1}\sum_{i=1}^{n}\phi_k\left(X_{ij}\right)$, $\psi_{jk}\left(x\right)=\phi_k\left(x\right)-\bar{\phi}_{jk}$, $Z_{ij}=\left(\psi_{j1}\left(X_{ij}\right),...,\psi_{jm_n}\left(X_{ij}\right)\right)'$, $\mathbf{Z}_j=\left(Z_{1j},...,Z_{nj}\right)'$, $\mathbf{Z}=(\mathbf{Z}_1,...,\mathbf{Z}_p)$, and $\mathbf{Y}=\left(Y_1-\bar{Y},...,Y_n-\bar{Y}\right)'$ where $\bar{Y}=n^{-1}\sum_{i=1}^{n}Y_i$. Note that $\mathbf{Z}_j$ is an $n\times m_n$ "design" matrix for the jth covariate. It is easy to verify that minimizing (3.3) subject to (3.4) is equivalent to minimizing

$$L_n\left(\boldsymbol{\beta}_n;\lambda_n\right)=\left\|\mathbf{Y}-\mathbf{Z}\boldsymbol{\beta}_n\right\|_2^2+\lambda_n\sum_{j=1}^{p}w_{nj}\left\|\boldsymbol{\beta}_{nj}\right\|_2. \tag{3.5}$$

In the first step, Huang, Horowitz, and Wei compute the group Lasso estimator by minimizing $L_{n1}(\boldsymbol{\beta}_n;\lambda_{n1}) = \|\mathbf{Y}-\mathbf{Z}\boldsymbol{\beta}_n\|_2^2 + \lambda_{n1}\sum_{j=1}^p \|\boldsymbol{\beta}_{nj}\|_2$, which is a special case of (3.5) by setting $w_{nj}=1$ for $j=1,...,p$ and $\lambda_n=\lambda_{n1}$. Denote the resulting group Lasso estimator as $\tilde{\boldsymbol{\beta}}_n \equiv \tilde{\boldsymbol{\beta}}_n(\lambda_{n1})$. In the second step they minimize the adaptive group Lasso objective function $L_n(\boldsymbol{\beta}_n;\lambda_{n2})$ by choosing

$$w_{nj} = \begin{cases} ||\tilde{\boldsymbol{\beta}}_{nj}||_2^{-1} & \text{if } ||\tilde{\boldsymbol{\beta}}_{nj}||_2 > 0, \\ \infty & \text{if } ||\tilde{\boldsymbol{\beta}}_{nj}||_2 = 0. \end{cases}$$

Denote the adaptive group Lasso estimator of $\boldsymbol{\beta}_n$ as $\hat{\boldsymbol{\beta}}_n \equiv \hat{\boldsymbol{\beta}}_n(\lambda_{n2}) = (\hat{\boldsymbol{\beta}}_{n1}', ..., \hat{\boldsymbol{\beta}}_{np}')'$. The adaptive group Lasso estimators of μ and f_j are then given by

$$\hat{\mu}_n = \bar{Y} \text{ and } \hat{f}_{nj}(x) = \sum_{k=1}^{m_n} \hat{\boldsymbol{\beta}}_{nj}\psi_k(x),\ 1 \le j \le p.$$

Assume $f_j(x) \neq 0$ for $j \in A_1 = \{1,...,p^*\}$, $=0$ for $j \in A_0 = \{p^*+1,...,p\}$. Let $\hat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n$ denote $\text{sgn}_0(||\hat{\boldsymbol{\beta}}_{nj}||) = \text{sgn}_0(||\boldsymbol{\beta}_{nj}||)$, $1 \le j \le p$, where $\text{sgn}_0(|x|) = 1$ if $|x|>0$ and $=0$ if $|x|=0$. Under some regularity conditions, Huang, Horowitz, and Wei show that $P(\hat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n) \to 1$ as $n \to \infty$, and $\sum_{j=1}^{p^*} ||\hat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}||_2^2 = O_P(m_n^2/n + 1/m_n^{2d-1} + 4m_n^2\lambda_{n2}^2/n^2)$, where d denotes the smoothness parameter of f_j's (e.g., $d=2$ if each f_j has continuous second order derivative). In terms of the estimators of the nonparametric components, they show that

$$P\left(\left\|\hat{f}_{nj}\right\|_2 > 0,\ j \in A_1 \text{ and } \left\|\hat{f}_{nj}\right\|_2 = 0,\ j \in A_0\right) \to 1 \text{ as } n \to \infty$$

and

$$\sum_{j=1}^{p^*} \left\|\hat{f}_{nj} - f_j\right\|_2^2 = O_P\left(\frac{m_n}{n} + \frac{1}{m_n^{2d}} + \frac{4m_n\lambda_{n2}^2}{n^2}\right).$$

The above result states that the adaptive group Lasso can consistently distinguish nonzero components from zero components, and gives an upper bound on the rate of convergence of the estimators.

3.2 Meier, Geer, and Bühlmann's (2009) sparsity-smoothness penalty

Meier, Geer, and Bühlmann (2009) consider the problem of estimating a high-dimensional additive model in (3.1) where $p = p_n \gg n$. For identification purpose, they assume that all f_j's are centered, i.e., $\sum_{i=1}^n f_j(X_{ij}) = 0$ for $j=1,...,p$. For any vector $a = (a_1,...,a_n)' \in \mathbb{R}^n$, define $\|a\|_n^2 \equiv n^{-1}\sum_{i=1}^n a_i^2$. Let $f_j = (f_j(X_{j1}),...,f_j(X_{jn}))'$. Meier, Geer, and Bühlmann define the sparsity-smoothness penalty as

$$J(f_j) = \lambda_{1n}\sqrt{\|f_j\|_n^2 + \lambda_{2n} I^2(f_j)}$$

where $I^2(f_j) = \int [f_j''(x)]^2 dx$ measures the smoothness of f_j with $f_j''(x)$ denoting the second order derivative of $f_j(x)$, and λ_{1n} and λ_{2n} are two tuning parameters controlling the amount of penalization. The estimator is given by the following PLS problem:

$$\left(\hat{f}_1,...,\hat{f}_p\right) = \underset{f_1,...,f_p \in \mathcal{F}}{\text{argmin}} \left\|Y - \sum_{j=1}^p f_j\right\|_n^2 + \sum_{j=1}^p J(f_j) \tag{3.6}$$

where $\mathcal{F}$ is a suitable class of functions and $Y = (Y_1, ..., Y_n)'$. They choose B-splines to approximate each function f_j: $f_j(x) = \sum_{k=1}^{m_n} \beta_{jk}\phi_{jk}(x)$, where $\phi_{jk}(x) : \mathbb{R} \to \mathbb{R}$ are the B-spline basis functions and $\boldsymbol{\beta}_j = \left(\beta_{j1}, ..., \beta_{jm_n}\right)' \in \mathbb{R}^{m_n}$ is the parameter vector corresponding to f_j. Let B_j denote the $n \times m_n$ design matrix of B-spline basis of the jth predictor. Denote the $n \times pm_n$ design matrix as $B = [B_1, B_2, ..., B_p]$. By assuming that all f_j's are second order continuously differentiable, one can reformulate (3.6) as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|Y - B\boldsymbol{\beta}\|_n^2 + \lambda_{1n} \sum_{j=1}^{p} \sqrt{\frac{1}{n}\boldsymbol{\beta}_j' B_j B_j' \boldsymbol{\beta}_j + \lambda_{2n}\boldsymbol{\beta}_j' \Omega_j \boldsymbol{\beta}_j} \tag{3.7}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_p')'$, and Ω_j is an $m_n \times m_n$ matrix with (k,l)th element given by $\Omega_{j,kl} = \int \phi_{jk}''(x)\,\phi_{jl}''(x)\,dx$ for $k, l \in \{1, ..., m_n\}$. Let $M_j = \frac{1}{n}B_j B_j' + \lambda_{2n}\Omega_j$. Then (3.7) can be written as a general group Lasso problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|Y - B\boldsymbol{\beta}\|_n^2 + \lambda_{1n} \sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_j' M_j \boldsymbol{\beta}_j}. \tag{3.8}$$

By Cholesky decomposition, $M_j = R_j' R_j$ for some $m_n \times m_n$ matrix R_j. Define $\tilde{\boldsymbol{\beta}}_j = R_j \boldsymbol{\beta}_j$ and $\tilde{B}_j = B_j R_j^{-1}$. (3.8) reduces to

$$\hat{\tilde{\boldsymbol{\beta}}} = \underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}} \left\| Y - \tilde{B}\tilde{\boldsymbol{\beta}} \right\|_n^2 + \lambda_{1n} \sum_{j=1}^{p} \left\| \tilde{\boldsymbol{\beta}}_j \right\| \tag{3.9}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', ..., \tilde{\boldsymbol{\beta}}_p')'$, and $||\tilde{\boldsymbol{\beta}}_j||$ denotes the Euclidean norm in $\mathbb{R}^{m_n}$. For fixed λ_{2n}, (3.9) is an ordinary group Lasso problem. For large enough λ_{1n} some of the coefficient group $\boldsymbol{\beta}_j \in \mathbb{R}^{m_n}$ will be estimated as exactly zero.

Meier, Geer, and Bühlmann argue empirically that the inclusion of a smoothness-part ($I^2(f_j)$) into the penalty functions yields much better results than having the sparsity-term ($\|f_j\|_n$) only. Under some conditions, the procedure can select a set of f_j's containing all the additive nonzero components. However, the model selection consistency of their procedure is not established. The selected set may include zero components and then be larger than the set of nonzero f_j's.

3.3 Ravikumar, Lafferty, Liu, and Wasserman's (2009) sparse additive models

Ravikumar, Lafferty, Liu, and Wasserman (2009) propose a new class of methods for high dimensional nonparametric regression and classification called *SParse Additive Models* (SPAM). The models combine ideas from sparse linear modelling and additive nonparametric regression. The models they consider take the form (3.1), but they restrict $X_i = (X_{i1}, ..., X_{ip})' \in [0,1]^p$ and $p = p_n$ can diverge with n. They consider a modification of standard additive model optimization problem as follows

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} E\left[Y_i - \sum_{j=1}^{p} \beta_j g_j(X_{ij}) \right]^2$$

subject to

$$\sum_{j=1}^{p} \left|\beta_j\right| \leq L \text{ and } E\left[g_j\left(X_{ij}\right)^2\right] = 1,\ j = 1, ..., p,$$

where $\mathcal{H}_j$ denotes the Hilbert subspace $L_2\left(\mu_j\right)$ of measurable function $f_j\left(X_{ij}\right)$ with $E\left[f_j\left(X_{ij}\right)\right] = 0$ and μ_j being the marginal distribution of X_{ij}, $\beta = (\beta_1, ..., \beta_p)'$, and β_j is the rescaling parameter such that $f_j = \beta_j g_j$ and $E[g_j\left(X_{ij}\right)^2] = 1$. The constraint that $\boldsymbol{\beta}$ lies in the l_1-ball $\{\beta : \|\beta\|_1 \leq L\}$ induces sparsity for the estimate $\hat{\beta}$ of β as for the Lasso estimate. Absorbing β_j in f_j, we can re-express the minimization problem in the equivalent Lagrangian form

$$L\left(f, \lambda\right) = \frac{1}{2} E\left[Y_i - \sum_{j=1}^{p} f_j\left(X_{ij}\right)\right]^2 + \lambda_n \sum_{j=1}^{p} \sqrt{E\left[f_j^2\left(X_{ij}\right)\right]}, \tag{3.10}$$

where λ_n is a tuning parameter. The minimizers for (3.10) satisfy

$$f_j\left(X_{ij}\right) = \left[1 - \frac{\lambda_n}{\sqrt{E\left[P_j\left(X_{ij}\right)^2\right]}}\right]_+ P_j\left(X_{ij}\right) \quad \text{almost surely (a.s.)},$$

where $[\cdot]_+$ denotes the positive part of $\cdot$, and $P_j\left(X_{ij}\right) = E\left(R_{ij}|X_{ij}\right)$ denotes the projection of the residuals $R_{ij} = Y_i - \sum_{k \neq j} f_k\left(X_{ik}\right)$ onto $\mathcal{H}_j$.

To get a sample version of the above solution, Ravikumar et al. insert the sample estimates into the population algorithm as in standard backfitting. The projection $P_j = \left(P_j\left(X_{1j}\right), ..., P_j\left(X_{nj}\right)\right)'$ can be estimated by smoothing the residuals:

$$\hat{P}_j = \mathcal{S}_j R_j$$

where $\mathcal{S}_j$ is a linear smoother (e.g., an $n \times n$ matrix for the local linear or sieve smoother) and $R_j = (R_{1j}, .., R_{nj})'$. Let $\hat{s}_j = \sqrt{n^{-1} \sum_{i=1}^{n} \hat{P}_{ij}^2}$ be an estimator of $s_j = \sqrt{E[P_j\left(X_{ij}\right)^2]}$ where $\hat{P}_{ij}$ denotes the ith element of $\hat{P}_j$. They propose the SPAM backfitting algorithm to solve f_j's as follows: Given regularization parameter λ, initialize $\hat{f}_j\left(X_{ij}\right) = 0$ for $j = 1, ..., p$, and then iterate the following steps until convergence, for each $j = 1, .., p$:

1. Compute the residual, $R_{ij} = Y_i - \sum_{k \neq j}^{p} \hat{f}_k\left(X_{ik}\right)$;
2. Estimate P_j by $\hat{P}_j = \mathcal{S}_j R_j$;
3. Estimate s_j by $\hat{s}_j$;
4. Obtain the soft thresholding estimate $\hat{f}_j\left(X_{ij}\right) = [1 - \lambda / \hat{s}_i]_+ \hat{P}_{ij}$;
5. Center $\hat{f}_j$ to obtain $\hat{f}_j\left(X_{ij}\right) - n^{-1} \sum_{i=1}^{n} \hat{f}_j\left(X_{ij}\right)$ and use this as an updated estimate of $f_j\left(X_{ij}\right)$.

The outputs are $\hat{f}_j$, based on which one can also obtain $\sum_{i=1}^{p} \hat{f}_j\left(X_{ij}\right)$.

If $f_j(x)$ can be written in terms of orthonormal basis functions $\{\psi_{jk} : k = 1, 2, ...\}$: $f_j(x) = \sum_{k=1}^{\infty} \beta_{jk}\psi_{jk}(x)$ with $\beta_{jk} = \int f_j(x)\psi_{jk}(x)dx$, we can approximate it by $\tilde{f}_j(x) = \sum_{k=1}^{m_n} \beta_{jk}\psi_{jk}(x)$ where $m_n \to \infty$ as $n \to \infty$. In this case, the smoother $\mathcal{S}_j$ can be taken to be the least squares projection onto the truncated set of basis $\{\psi_{j1}, ..., \psi_{jm_n}\}$. The orthogonal series smoother is $\mathcal{S}_j = \Psi_j(\Psi_j'\Psi_j)^{-1}\Psi_j'$, where Ψ_j denotes the $n \times m_n$ design matrix with (i,k)th element given by $\psi_{jk}(X_{ij})$. Then the backfitting algorithm reduces to choosing $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_p')'$ to minimize

$$\frac{1}{2n}\left\|Y - \sum_{j=1}^{p}\Psi_j\boldsymbol{\beta}_j\right\|_2^2 + \lambda_n\sum_{j=1}^{p}\sqrt{\frac{1}{n}\boldsymbol{\beta}_j'\left(\Psi_j'\Psi_j\right)\boldsymbol{\beta}_j}, \tag{3.11}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jm_n})'$. Note (3.11) is a sample version of (3.10) and it can be regarded as a functional version of the group Lasso by using the similar transformation form as used to obtain (3.9). Combined with the soft thresholding step, the update of f_j in the above algorithm can be thought as to minimize

$$\frac{1}{2n}\left\|R_j - \Psi_j\boldsymbol{\beta}_j\right\|_2^2 + \lambda_n\sqrt{\frac{1}{n}\boldsymbol{\beta}_j'\left(\Psi_j'\Psi_j\right)\boldsymbol{\beta}_j}.$$

Under some strong conditions, they show that with truncated orthogonal basis the SPAM backfitting algorithm can recover the correct sparsity pattern asymptotically if the number of relevant variables p^* is bounded. [p^* denotes the cardinality of the set $\{1 \le j \le p : f_j \ne 0\}$]. That is, their estimator can achieve the selection consistency.

3.4 Xue's (2009) SCAD procedure

Xue (2009) considers a penalized polynomial spline method for simultaneous model estimation and variable selection in the additive model (3.1) where p is fixed. Let $\mathcal{M}_n = \{m_n(x) = \sum_{l=1}^{p} g_l(x_l) : g_l \in \varphi_l^{0,n}\}$ be the approximation space, where $\varphi_l^{0,n} = \{g_l : n^{-1}\sum_{i=1}^{n} g_l(X_{il}) = 0,\ g_l \in \varphi_l\}$ and φ_l is the space of empirically centered polynomial splines of degree $q \ge 1$ on the lth intervals constructed by interior knots on [0,1]. The penalized least squares estimator is given by

$$\hat{m} = \underset{m_n = \sum_{l=1}^{p} f_l \in \mathcal{M}_n}{\operatorname{argmin}}\left[\frac{1}{2}\|Y - m_n\|_n^2 + \sum_{l=1}^{p} p_{\lambda_n}(\|f_l\|_n)\right]$$

where $Y = (Y_1, ..., Y_n)'$, $m_n = (m_n(X_1), ..., m_n(X_n))'$, and $p_{\lambda_n}(\cdot)$ is a given penalty function depending on a tuning parameter λ_n. Different penalty functions lead to different variable selection procedures. Noting the desirable properties of the SCAD, they use the spline SCAD penalty.

The proposed polynomial splines have polynomial spline basis representation. Let $J_l = N_l + q$ and $B_l = \{B_{l1}, ..., B_{lJ_l}\}$ be a basis for $\varphi_l^{0,n}$. For any fixed $x = (x_1, ..., x_p)'$, let $B_l(x_l) = [B_{l1}(x_l), ..., B_{lJ_l}(x_l)]'$. One can express $\hat{m}$ as

$$\hat{m}(x) = \sum_{l=1}^{p}\hat{f}_l(x_l) \text{ and } \hat{f}_l(x_l) = \hat{\boldsymbol{\beta}}_l'B_l(x_l) \text{ for } l = 1, ..., p,$$

where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', ..., \hat{\boldsymbol{\beta}}_p')'$ minimizes the PLS criterion:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}=(\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_d')'}{\operatorname{argmin}} \left[\frac{1}{2} \left\| Y - \sum_{l=1}^{p} \boldsymbol{\beta}_l' B_l \right\|_n^2 + \sum_{l=1}^{p} p_{\lambda_n} \left(\|\boldsymbol{\beta}_l\|_{K_l} \right) \right]$$

where $\|\boldsymbol{\beta}_l\|_{K_l} = \sqrt{\boldsymbol{\beta}_l' K_l \boldsymbol{\beta}_l}$ with $K_l = n^{-1} \sum_{i=1}^{n} B_l(X_{il}) B_l(X_{il})'$.

Under some mild conditions, she shows that the SCAD penalized procedure estimates the non-zero function components with the same optimal mean square convergence rate as the standard polynomial spline estimators, and correctly sets the zero function components to zero with probability approaching one as the sample size n goes to infinity.

4 Variable Selection in Partially Linear Models

In this section, we consider the problem of variable selection in the following *partially linear model* (PLM)

$$Y_i = \beta' X_i + g(Z_i) + \varepsilon_i, \ i = 1, 2, ..., n, \tag{4.1}$$

where $X_i = (X_{i1}, ..., X_{ip})'$ is a $p \times 1$ vector of regressors that enter the model linearly, Z_i is a $q \times 1$ vector of regressors that enter the model with an unknown functional form g, and ε_i is an error term such that

$$E(\varepsilon_i | X_i, Z_i) = 0 \text{ a.s.} \tag{4.2}$$

To allow p to increase as n increases, we sometimes write p as p_n below.

Xie and Huang (2009) consider the problem of simultaneous variable selection and estimation (4.1) with a divergent number of covariates in the linear part. Ni, Zhang, and Zhang (2009) propose a *double-penalized least squares* (DPLS) approach to simultaneously achieve the estimation of the nonparametric component g and the selection of important variables in X_i in (4.1). Kato and Shiohama (2009) consider variable selection in (4.1) in the time series framework. In the large p framework, Chen, Yu, Zou, and Liang (2012) propose to use the adaptive Elastic-net for variable selection for parametric components by using profile least squares approach to convert the partially linear model to a classical linear regression model. Liang and Li (2009) consider variable selection in the PLM (4.1) where Z_i is a scalar random variable but X_i is measured with error and is not observable. In addition, Liu, Wang, and Liang (2011) consider the additive PLM where $g(Z_i) = \sum_{k=1}^{q} g_k(Z_{ik})$ in (4.1). Besides, it is worth mentioning that Bunea (2004) considers covariate selection in X_i when the dimension of X_i is fixed and Z_i is a scalar variable in (4.1) based on a BIC type of information criterion and a sieve approximation for $g(\cdot)$. He shows that one can consistently estimate the subset of nonzero coefficients of the linear part and establish its oralce property. But we will not review this paper in detail as the procedure is not a simultaneous variable selection and estimation procedure.

4.1 Xie and Huang's (2009) SCAD-penalized regression in high-dimension PLM

Xie and Huang (2009) consider the problem of simultaneous variable selection and estimation in (4.1) when $p = p_n$ is divergent with n and q is fixed. To make it explicit that the coefficients depend on n, one can write $\beta = \beta^{(n)}$. They allow the number p_n^* of nonzero components in $\beta^{(n)}$ diverge with n too.

Since g is unknown, Xie and Huang use the polynomial splines to approximate it. Let $\{B_{nw}(z) : 1 \leq w \leq m_n\}$ be a sequence of basis functions. Let $B(z) = (B_{n1}(z), ..., B_{nm_n}(z))'$, $\mathbf{B}^{(n)}$ be the $n \times m_n$ matrix whose ith row is $B(Z_i)'$. Under some smoothness conditions, $g(z)$ can be well approximated by $\alpha^{(n)\prime} B(z)$ for some $\alpha^{(n)} \in \mathbb{R}^{m_n}$. Then the problem of estimating g becomes that of estimating $\alpha^{(n)}$. Let $Y = (Y_1, ..., Y_n)'$ and $\mathbf{X}^{(n)} = (X_1, ..., X_n)'$. Then one can consider the following PLS objective function for estimating $\beta^{(n)}$ and $\alpha^{(n)}$ with the SCAD penalty

$$Q_n\left(\beta^{(n)}, \alpha^{(n)}\right) = \left\| Y - \mathbf{X}^{(n)}\beta^{(n)} - \mathbf{B}^{(n)}\alpha^{(n)} \right\|_2^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}\left(\left|\beta_j^{(n)}\right|\right)$$

where $\beta_j^{(n)}$ denotes the jth element of $\beta^{(n)}$, and $p_{\lambda_n}(\cdot)$ is the SCAD penalty function with λ_n as a tuning parameter. Let $(\hat{\beta}^{(n)}, \hat{\alpha}^{(n)})$ denote the solution to the above minimization problem, and $\hat{g}_n(z) = \hat{\alpha}^{(n)\prime} B(z)$.

For any $\beta^{(n)}$, $\alpha^{(n)}$ minimizing Q_n satisfies that

$$\mathbf{B}^{(n)\prime}\mathbf{B}^{(n)}\alpha^{(n)} = \mathbf{B}^{(n)\prime}(Y - \mathbf{B}^{(n)}\beta^{(n)}).$$

Let $P_{\mathbf{B}^{(n)}} = \mathbf{B}^{(n)}(\mathbf{B}^{(n)\prime}\mathbf{B}^{(n)})^{-1}\mathbf{B}^{(n)\prime}$ be the projection matrix. It is easy to verify that the profile least squares objective function of the parametric part becomes

$$\tilde{Q}_n\left(\beta^{(n)}\right) = \left\| (I - P_{\mathbf{B}^{(n)}})\left(Y - \mathbf{X}^{(n)}\beta^{(n)}\right)\right\|^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}\left(\left|\beta_j^{(n)}\right|\right),$$

and $\hat{\beta}^{(n)} = \text{argmin}_{\beta^{(n)}} \tilde{Q}\left(\beta^{(n)}\right)$.

Under some regularity conditions that allow divergent p_n^*, Xie and Huang show that variable selection is consistent, and the SCAD penalized estimators of the nonzero coefficients possess the oracle properties, and the estimator of the nonparametric estimate can achieve the optimal convergence rate.

4.2 Ni, Zhang, and Zhang's (2009) double-penalized least squares regression in PLM

Ni, Zhang, and Zhang (2009) consider a unified procedure for variable selection in the PLM in (4.1) when Z_i is restricted to be a scalar variable on $[0, 1]$. To simultaneously achieve the estimation of the nonparametric component g and the selection of important variables in X_i, they propose a *double-penalized least squares* (DPLS) approach by minimizing

$$Q(\beta, g) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \beta' X_i - g(Z_i)\right]^2 + \frac{n\lambda_{1n}}{2}\int_0^1 \left[g''(z)\right]^2 dz + n\sum_{j=1}^{p} p_{\lambda_{2n}}\left(|\beta_j|\right). \tag{4.3}$$

The first penalty term in (4.3) penalizes the roughness of the nonparametric fit $g(z)$ and the second penalty term imposes the usual SCAD penalty on the finite dimensional parameter β. Let $\mathbf{X} = (X_1, ..., X_n)'$, $Y = (Y_1, ..., Y_n)'$, and $\mathbf{g} = (g(Z_1), ..., g(Z_n))'$. It can be shown that the given λ_{1n} and λ_{2n}, minimizing (4.3) leads to a smoothing spline estimate for g and one can rewrite the DPLS (4.3) as

$$Q_{dp}(\beta, g) = \frac{1}{2}(Y - \mathbf{X}\beta - \mathbf{g})'(Y - \mathbf{X}\beta - \mathbf{g}) + \frac{n\lambda_{1n}}{2}\mathbf{g}'\mathbf{K}\mathbf{g} + n\sum_{j=1}^{p} p_{\lambda_{2n}}(|\beta_j|), \tag{4.4}$$

where $\mathbf{K}$ is the nonnegative definite smoothing matrix defined by Green and Silverman (1994). Given β, one can obtain the minimizer of $\mathbf{g}$ as $\hat{\mathbf{g}}(\beta) = (I_n + n\lambda_{1n}\mathbf{K})^{-1}(Y - \mathbf{X}\beta)$ where I_n is an $n \times n$ identity matrix. With this, one can obtain readily the profile PLS objective function of β as follows

$$Q(\beta) = \frac{1}{2}(Y - \mathbf{X}\beta)'\left[I_n - (I_n + n\lambda_{1n}\mathbf{K})^{-1}\right](Y - \mathbf{X}\beta) + n\sum_{j=1}^{p} p_{\lambda_{2n}}(|\beta_j|).$$

Let $\hat{\beta}$ denote the minimizing solution to the above problem. Ni, Zhang, and Zhang show that $\hat{\beta}$ has the oracle properties in the case of fixed p under some regularity conditions. In the case where $p = p_n$ is divergent with $p_n \ll n$, they also establish the selection consistency by allowing the number p_n^* of nonzero components in β to be divergent at a slow rate.

4.3 Kato and Shiohama's (2009) partially linear models

Kato and Shiohama (2009) consider the PLM in (4.1) in the time series framework by restricting $Z_i = t_i = i/n$ and allowing ε_i to be a linear process. They assume that $g(t_i)$ is an unknown time trend function that can be *exactly* expressed as

$$g(t_i) = \sum_{k=1}^{m} \alpha_k \phi_k(t_i) = \alpha' \boldsymbol{\phi}_i$$

where $\boldsymbol{\phi}_i = (\phi_1(t_i), ..., \phi_m(t_i))'$ is an m-dimensional vector constructed from basis functions $\{\phi_k(t_i) : k = 1, ..., m\}$, and $\alpha = (\alpha_1, ..., \alpha_m)'$ is an unknown parameter vector to be estimated. They propose variable selection via the PLS method:

$$\|Y - X\beta - \boldsymbol{\phi}\alpha\|_2^2 + n\lambda_{0n}\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} + n\left(\sum_{j=1}^{p} p_{\lambda_{1n}}(|\beta_j|) + \sum_{k=1}^{m} p_{\lambda_{2n}}(|\alpha_k|)\right)$$

where $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_n)'$, λ_{0n} in the second term is used to control the trade-off between the goodness-of-fit and the roughness of the estimated function, $\mathbf{K}$ is an appropriate positive semi-definite symmetric matrix, $p_{\lambda_{in}}(\cdot)$ are penalty functions and $\lambda_{in}, i = 1, 2$, are regularization parameters, which control the model complexity. They consider several different penalty functions: the hard thresholding penalty, l_2-penalty in ridge regression, the Lasso penalty, and the SCAD penalty. Under some conditions, they establish the convergence rates for the PLS estimator, and show its oracle property.

4.4 Chen, Yu, Zou, and Liang's (2012) adaptive Elastic-Net estimator

Chen, Yu, Zou, and Liang (2012) propose to use the adaptive Elastic-net (Zou and Zhang, 2009) for variable selection for parametric components when the dimension p is large, using profile least squares approach to convert the PLM to a classical linear regression model.

Noting that $E(Y_i|Z_i) = \beta' E(X_i|Z_i) + g(Z_i)$ under (4.1)-(4.2), we have

$$Y_i - E(Y_i|Z_i) = \beta' [X_i - E(X_i|Z_i)] + \varepsilon_i, \tag{4.5}$$

which is a standard linear model if $E(Y_i|Z_i)$ and $E(X_i|Z_i)$ were known. Let $\hat{E}(Y_i|Z_i)$ and $\hat{E}(X_i|Z_i)$ be the local linear estimators for $E(Y_i|Z_i)$ and $E(X_i|Z_i)$, respectively. Let $\hat{X}_i = X_i - \hat{E}(X_i|Z_i)$, $\hat{Y}_i = Y_i - \hat{E}(Y_i|Z_i)$, $\hat{X} = (\hat{X}_1, ..., \hat{X}_n)'$, and $\hat{Y} = (\hat{Y}_1, ..., \hat{Y}_n)'$. Chen, Yu, Zou, and Liang's adaptive Elastic-Net procedure is composed of the following two steps:

1. Construct the Elastic-Net estimator of β given by

$$\hat{\beta}_{\text{Enet}} = \left(1 + \frac{\lambda_2}{n}\right) \underset{\beta}{\text{argmin}} \left\{ \left\| \hat{Y} - \hat{X}\beta \right\|_2^2 + \lambda_{2n} \|\beta\|_2^2 + \lambda_{1n} \|\beta\|_1 \right\}.$$

2. Let $\hat{w}_j = |\hat{\beta}_{\text{Enet},j}|^{-\gamma}$ for $j = 1, ..., p$ and some $\gamma > 0$, where $\hat{\beta}_{\text{Enet},j}$ denotes the jth element of $\hat{\beta}_{\text{Enet}}$. The adaptive Elastic-Net estimator of β is given by

$$\hat{\beta}_{\text{AdaEnet}} = \left(1 + \frac{\lambda_2}{n}\right) \underset{\beta}{\text{argmin}} \left\{ \left\| \hat{Y} - \hat{X}\beta \right\|_2^2 + \lambda_{2n} \|\beta\|_2^2 + \lambda_{1n}^* \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right\}.$$

Here, the l_1 regularization parameters λ_{1n} and λ_{1n}^* control the sparsity of the Elastic-Net and adaptive Elastic-Net estimators, respectively. The same λ_{2n} for the l_2-penalty is used in both steps. Under some regular conditions that allows diverging p, they show that profiled adaptive Elastic-Net procedure has the oracle property. In particular, $P(\{j : \hat{\beta}_{\text{AdaEnet},j} \neq 0\} = \mathcal{A}) \to 1$ as $n \to \infty$ where $\mathcal{A} = \{j : \beta_j \neq 0, j = 1, ..., p\}$.

4.5 Liang and Li's (2009) variable selection with measurement errors

Liang and Li (2009) also consider the model in (4.1)-(4.2) where Z_i is a scalar random variable, and X_i is measured with error and is not observable. Let W_i denote the observed surrogate of X_i, i.e.,

$$W_i = X_i + U_i \tag{4.6}$$

where U_i is the measurement error with mean zero and unknown covariance Σ_{uu}. Assume U_i is independent of (X_i, Z_i, Y_i). Since $E(U_i|Z_i) = 0$, $g(Z_i) = E(Y_i|Z_i) - E(W_i|Z_i)' \beta$. The PLS function based on partial residuals is defined as

$$L_p(\Sigma_{uu}, \beta) = \frac{1}{2} \sum_{i=1}^{n} \left\{ [Y_i - \hat{m}_y(Z_i)] - [W_i - \hat{m}_w(Z_i)]' \beta \right\} - \frac{n}{2} \beta' \Sigma_{uu} \beta + n \sum_{j=1}^{p} q_{\lambda_j}(|\beta_j|), \tag{4.7}$$

where $\hat{m}_y(Z_i)$ and $\hat{m}_w(Z_i)$ are estimators of $E(Y_i|Z_i)$ and $E(W_i|Z_i)$, respectively, and $q_{\lambda_j}(\cdot)$ is a penalty function with a tuning parameter λ_j. The second term is included to correct the bias in the squared loss function due to the presence of measurement error.

The PLS function (4.7) provides a general framework of variable selection in PLMs with measurement errors. In principle, one can use all kinds of penalty functions. But Liang and Li focus on the case of SCAD penalty. Under some conditions, they show that with probability approaching one, there exists a $\sqrt{n}$-consistent PLS estimator $\hat{\beta}$ when Σ_{uu} is estimated from partially replicated observations, and the estimator $\hat{\beta}$ possesses the oracle properties. In addition, they also consider the penalized quantile regression for PLMs with measurement error. See Section 8.1.

4.6 Liu, Wang, and Liang's (2011) additive PLMs

Liu, Wang, and Liang (2011) consider the additive PLM of this form

$$Y_i = \beta' X_i + \sum_{k=1}^{q} g_k(Z_{ik}) + \varepsilon_i$$

where $X_i = (X_{i1}, ..., X_{ip})'$ and $Z_i = (Z_{i1}, ..., Z_{iq})'$ enter the linear and nonparametric components, respectively, $g_1, ..., g_q$ are unknown smooth functions, and $E(\varepsilon_i|X_i, Z_i) = 0$ a.s. They are interested in the variable selection in the parametric component. For identification, assume that $E[g_k(Z_{ik})] = 0$ a.s. for $k = 1, ..., q$. In addition, they assume that both p and q are fixed.

Since g_k's are unknown, Liu, Wang, and Liang propose to use spline approximation. Let $\mathcal{S}_n$ be the space of polynomial functions on $[0, 1]$ of degree $\varrho \geq 1$. Let $\mathcal{G}_n$ be the collection of functions g with additive form $g(z) = g_1(z_1) + \cdots + g_K(z_K)$. For the kth covariate z_k, let $b_{jk}(z_k)$ be the B-spline basis functions of degree ϱ. For any $g \in \mathcal{G}_n$, one can write

$$g(z) = \gamma' b(z)$$

where $b(z) = \{b_{jk}(z_k), j = -\varrho, ..., m_n,\ k = 1, ..., q\}' \in \mathbb{R}^{m_n q}$, and γ is the corresponding vector of coefficients and its elements are arranged in the same order as $b(z)$. The PLS objective function is given by

$$L_P(\beta, \gamma) = \frac{1}{2} \sum_{i=1}^{n} \left[Y_i - \gamma' b(Z_i) - \beta' X_i\right]^2 + n \sum_{j=1}^{p} q_{\lambda_j}\left(|\beta_j|\right),$$

where $q_{\lambda_j}(\cdot)$ is a penalty function with a tuning parameter λ_j. Let $\hat{\beta}$ be the PLS estimator. Liu, Wang, and Liang consider the SCAD penalty function and show that the SCAD variable selection procedure can effectively identify the significant components with the associated parametric estimators satisfying the oracle properties. But they do not study the asymptotic properties of the estimators of nonparametric components.

5 Variable Selection in Functional/Varying Coefficients Models

Nonparametric varying or functional coefficient models (VCMs or FCMs) are useful for studying the time-dependent or the variable-dependent effects of variables. Many methods have been proposed for estimation of these models. See, e.g., Fan and Zhang (1998) for local polynomial smoothing method and Huang, Wu, and Zhou (2002) and Qu and Li (2006) for basis expansion and spline method. Several procedures have been developed for variable selection and estimation simultaneously for these models. Wang and Xia (2009) propose adaptive group Lasso for variable selections in VCMs with fixed p based on kernel estimation; Lian (2010) extends their approach by using double adaptive lasso and allowing p to be divergent with n. Zhao and Xue (2011) consider SCAD variable selection for VCMs with measurement error. Li and Liang (2008) considers variable selection in generalized varying-coefficient partially linear models by using the SCAD penalty. In addition Wang, Chen, and Li (2007), Wang, Li, and Huang (2008) and Wei, Huang, and Li (2011) consider sieve-estimation-based variable selection in VCMs with longitudinal data where the penalty takes either the SCAD or group Lasso form.

5.1 Wang and Xia's (2009) kernel estimation with adaptive group Lasso penalty

Wang and Xia (2009) consider the following varying coefficient model (VCM)

$$Y_i = X_i'\beta(Z_i) + \varepsilon_i \tag{5.1}$$

where $X_i = (X_{i1}, ..., X_{ip})'$ is a $p \times 1$ vector of covariates, Z_i is a scalar variable that takes values on [0,1], and ε_i is the error term satisfying $E(\varepsilon_i | X_i, Z_i) = 0$ a.s. The coefficient vector $\beta(z) = (\beta_1(z), ..., \beta_p(z))' \in \mathbb{R}^p$ is an unknown but smooth function in z, whose true value is given by $\beta_0(z) = (\beta_{01}(z), ..., \beta_{0d}(z))'$. Without loss of generality, assume that there exists an integer $p^* \leq p$ such that $0 < E[\beta_{0j}^2(Z_i)] < \infty$ for any $j \leq p^*$ but $E[\beta_{0j}^2(Z_i)] = 0$ for $j > p^*$. The main objection is to select the variables in X_i with nonzero coefficients when p is fixed.

Let $B = (\beta(Z_1), ..., \beta(Z_n))'$ and $B_0 = (\beta_0(Z_1), ..., \beta_0(Z_n))'$. Note that the last $(p - p^*)$ columns for B_0 should be 0. The selection of variables becomes identifying sparse columns in the matrix B_0. Following the group Lasso of Yuan and Lin (2006), Wang and Xia propose the following PLS estimate

$$\hat{B}_\lambda = (\hat{\beta}_\lambda(Z_1), ..., \hat{\beta}_\lambda(Z_n))' = \text{argmin}_{B \in \mathbb{R}^{n \times p}} Q_\lambda(B)$$

where

$$Q_\lambda(B) = \sum_{i=1}^{n} \sum_{t=1}^{n} [Y_i - X_i\beta(Z_t)]^2 K_h(Z_t - Z_i) + \sum_{j=1}^{p} \lambda_j \|b_j\|, \tag{5.2}$$

$\lambda = (\lambda_1, ..., \lambda_p)'$ is a vector of tuning parameters, $K_h(z) = h^{-1}K(z/h)$, $K(\cdot)$ is a kernel function, h is a bandwidth parameter, b_j denotes the jth column of B for $j = 1, ..., p$, and $\|\cdot\|$ denotes the usual Euclidean norm. Let $\hat{b}_{\lambda,k}$ denote the kth column of $\hat{B}_\lambda$ for $k = 1, ..., p$ so that we can also

write $\hat{B}_\lambda = (\hat{b}_{\lambda,1}, ..., \hat{b}_{\lambda,p})$.They propose an iterated algorithm based on the idea of the local quadratic approximation of Fan and Li (2001). Let $\tilde{B}$ be an initial estimate of B_0 and $\hat{B}_\lambda^{(m)} = (\hat{b}_{\lambda,1}^{(m)}, ..., \hat{b}_{\lambda,p}^{(m)}) = (\hat{\beta}_\lambda^{(m)}(Z_1), ..., \hat{\beta}_\lambda^{(m)}(Z_n))'$ be the Lasso estimate in the mth iteration. The objective function in (5.2) can be locally approximated by

$$\sum_{i=1}^{n}\sum_{t=1}^{n}\left[Y_i - X_i\beta(Z_t)\right]^2 K_h(Z_t - Z_i) + \sum_{j=1}^{d}\lambda_j \frac{\|b_j\|^2}{\left\|\hat{b}_{\lambda,j}^{(m)}\right\|}.$$

Let $\hat{B}_\lambda^{(m+1)}$ denote the minimizer. Its ith row is given by

$$\hat{\beta}_\lambda^{(m+1)}(Z_t) = \left[\sum_{i=1}^{n} X_i X_i' K_h(Z_t - Z_i) + D^{(m)}\right]^{-1} \sum_{i=1}^{n} X_i Y_i K_h(Z_t - Z_i)$$

where $D^{(m)}$ is a $p \times p$ diagonal matrix with its jth diagonal component given by $\lambda_j / ||\hat{b}_{\lambda,j}^{(m)}||$, $j = 1, ..., p$. The estimate for $\beta(z)$ is

$$\hat{\beta}_\lambda(z) = \left[\sum_{i=1}^{n} X_i X_i' K_h(z - Z_i) + D^{(m)}\right]^{-1} \sum_{i=1}^{n} X_i Y_i K_h(z - Z_i).$$

Let $\hat{\beta}_{a,\lambda}(z) = (\hat{\beta}_{\lambda,1}(z), ..., \hat{\beta}_{\lambda,p^*}(z))'$ and $\hat{\beta}_{b,\lambda}(z) = (\hat{\beta}_{\lambda,p^*+1}(z), ..., \hat{\beta}_{\lambda,p}(z))'$. Under some regular conditions, Wang and Xia show that: (i) $P(\sup_{z\in[0,1]} ||\hat{\beta}_{\lambda,b}(z)|| = 0) \to 1$; and (ii) $\sup_{z\in[0,1]} ||\hat{\beta}_{a,\lambda}(z) - \hat{\beta}_{ora}(z)|| = o_P\left(n^{-2/5}\right)$, where

$$\hat{\beta}_{ora}(z) = \left[\sum_{i=1}^{n} X_{i(a)} X_{i(a)}' K_h(Z_t - Z_i)\right]^{-1} \sum_{i=1}^{n} X_{i(a)} Y_i K_h(z - Z_i)$$

stands for the oracle estimator, and $X_{i(a)} = (X_{i1}, ..., X_{ip_0})'$. The second part implies that $\hat{\beta}_{a,\lambda}(z)$ has the oracle property.

They also propose to choose the tuning parameters λ by

$$\lambda_j = \frac{\lambda_0}{n^{-1/2}||\tilde{\beta}_j||},$$

where $\tilde{\beta}_j$ is the jth column of the unpenalized estimate $\tilde{B}$. λ_0 can be selected according to the following BIC-type criterion

$$\text{BIC}_\lambda = \log(RRS_\lambda) + df_\lambda \times \frac{\log(nh)}{nh}$$

where $0 \le df_\lambda \le p$ is the number of nonzero coefficients identified by $\hat{B}_\lambda$ and RRS_λ is defined as

$$RRS_\lambda = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n}\left[Y_i - X_i'\hat{\beta}_\lambda(Z_t)\right]^2 K_h(Z_i - Z_t).$$

Under some conditions, they show that the tuning parameter $\hat{\lambda}$ selected by the BIC-type criterion can identify the true model consistently.

5.2 Lian's (2010) double adaptive group Lasso in high-dimensional VCMs

Lian (2010) studies the problem of simultaneous variable selection and constant coefficient identification in high-dimensional VCMs based on B-spline basis expansion. He considers the VCM in (5.1) but allows for $p = p_n \gg n$. In addition, he explicitly allows some nonzero coefficients in $\beta(Z_t)$ to be constant a.s.

Using spline expansions, $\beta(z)$ can be approximated by $\sum_{k=1}^{m_n} b_{jk} B_k(z)$, where $\{B_k(z)\}_{k=1}^{m_n}$ is a normalized B-spline basis. Lian proposes the following PLS estimate

$$\hat{b} = \text{argmin}_b \frac{1}{2} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{p} \sum_{k=1}^{m_n} X_{ij} b_{jk} B_k(Z_i) \right]^2 + n\lambda_1 \sum_{j=1}^{p} w_{1j} \|b_j\| + n\lambda_2 \sum_{j=1}^{p} w_{2j} \|b_j\|_c,$$

where λ_1, λ_2 are regularization parameters, $w_1 = (w_{11}, ..., w_{1p})'$ and $w_2 = (w_{21}, ..., w_{2p})'$ are two given vectors of weights, $b_j = (b_{j1}, ..., b_{jm_n})'$, $\|b_j\| = \sqrt{\sum_{k=1}^{m_n} b_{jk}^2}$, and $\|b_j\|_c = \sqrt{\sum_{k=1}^{m_n} \left[b_{jk} - \bar{b}_j \right]^2}$ with $\bar{b}_j = m_n^{-1} \sum_{k=1}^{m_n} b_{jk}$. The first penalty is used for identifying the zero coefficients while the second is used for identifying the nonzero constant coefficients.

The minimization problem can be solved by the locally quadratic approximation as Fan and Li (2001) and Wang and Xia (2009). He also proposes a BIC-type criterion to select λ_1 and λ_2. Under some suitable conditions, he shows that consistency in terms of both variable selection and constant coefficients identification can be achieved, and the oracle property of the constant coefficients can be established.

5.3 Zhao and Xue's (2011) SCAD variable selection for VCMs with measurement errors

Zhao and Xue (2011) consider variable selection for the VCM in (5.1) when the covariate X_i is measured with errors and Z_i is error-free. That is, X_i is not observed but measured with additive error:

$$\xi_i = X_i + V_i$$

where V_i is the measurement error that is assumed to be independent of $(X_i, Z_i, \varepsilon_i)$ and have zero mean and variance-covariance matrix Σ_{VV}.

Like Lian (2010), Zhao and Xue propose to approximate $\beta(z)$ by a linear combination of B-spline basis functions. Let $B(z) = (B_1(z), ..., B_{m_n}(z))'$, $W_i = (X_{i1} B(Z_i)', ..., X_{ip} B(Z_i)') = [I_p \otimes B(Z_i)] X_i$, and $b = (b_1', ..., b_p')'$. Let $\tilde{W}_i = (\xi_{i1} B(Z_i)', ..., \xi_{ip} B(Z_i)') = [I_p \otimes B(Z_i)] \xi_i$, where ξ_{ij} denotes the jth element in ξ_i. Observing that

$$E\left[\tilde{W}_i \tilde{W}_i' | X_i, Z_i \right] = W_i W_i' + \Omega(Z_i)$$

where $\Omega(Z_i) = [I_p \otimes B(Z_i)] \Sigma_{VV} [I_p \otimes B(Z_i)]'$, they propose a bias-corrected PLS objective function

$$Q(b) = \sum_{i=1}^{n} \left[Y_i - b' \tilde{W}_i \right]^2 - \sum_{i=1}^{n} \sum_{j=1}^{p} b' \Omega(Z_i) b + n \sum_{j=1}^{p} q_\lambda \left(\|b_j\|_H \right),$$

where $q_\lambda(\cdot)$ is the SCAD penalty function with λ as a tuning parameter and $\|b_j\|_H = (b'Hb)^{1/2}$ with $H = \int_0^1 B(z) B(z)' dz$. They establish the consistency of the variable selection procedure and derive the optimal convergence rate of the regularized estimators.

5.4 Li and Liang's (2008) variable selection in generalized varying-coefficient partially linear model

Li and Liang (2008) consider the generalized varying-coefficient partially linear model (GVCPLM)

$$g\{\mu(u,x,z)\} = x'\alpha(u) + z'\beta$$

where $\mu(u,x,z) = E(Y_i|U_i = u, X_i = x, Z_i = z)$, X_i is p-dimensional, Z_i is q-dimensional, and U_i is a scalar random variable. They assume that p and q are fixed and focus on the selection of significant variables in the parametric component based on i.i.d. observations (Y_i, U_i, X_i, Z_i), $i = 1, ..., n$. The conditional quasi-likelihood of Y_i is $Q(\mu(U_i, X_i, Z_i))$, where

$$Q(\mu, y) = \int_\mu^y \frac{s-y}{V(s)} ds$$

and $V(s)$ is a specific variance function. Then the penalized likelihood function is defined by

$$L(\alpha, \beta) = \sum_{i=1}^n Q\left[g^{-1}(X_i'\alpha(U_i) + Z_i'\beta), Y_i\right] - n\sum_{j=1}^q p_{\lambda_j}(|\beta_j|) \tag{5.3}$$

where β_j denotes the jth element of β, $p_{\lambda_j}(\cdot)$ is a specific penalty function with a tuning parameter λ_j, and g^{-1} denotes the inverse function of g. They propose to use the SCAD penalty. Since $\alpha(u)$ is unknown, they first use local likelihood technique to estimate $\alpha(u)$, then substitute the resulting estimate into the above penalized likelihood function and finally maximize (5.3) with respect to β. Under some conditions, they establish the rate of convergence for the resulting PLS estimator $\hat{\beta}$ of β. With proper choices of penalty functions and tuning parameters, they show the asymptotic normality of $\hat{\beta}$ and demonstrate that the proposed procedure performs as well as an oracle procedure. To select variables in X_i that is associated with the nonparametric component, they propose a *generalized likelihood ratio* (GLR) test statistic to test the null hypothesis of some selected components being zero.

5.5 Sieve-estimation-based variable selection in VCMs with longitudinal data

Several papers address the issue of variable selection in VCMs with longitudinal data. They base the variable selection on sieve estimation with either SCAD or Lasso penalty with balanced or unbalanced data.

Let $Y_i(t_j)$ be the expression level of the ith individual at time t_j where $i = 1, ..., n$ and $j = 1, ..., T$. Wang, Chen, and Li (2007) consider the following VCM

$$Y_i(t) = \mu(t) + \sum_{k=1}^p \beta_k(t) X_{ik} + \varepsilon_i(t) \tag{5.4}$$

where $\mu(t)$ indicates the overall mean effect, $\varepsilon_i(t)$ is the error term, and other objects are defined as above. They approximate $\beta_k(t)$ by using the natural cubic B-spline basis: $\beta_k(t) \approx \sum_{l=1}^{m_n} \beta_{kl} B_l(t)$ where $B_l(t)$ is the natural cubic B-spline basis function, for $l = 1, ..., m_n$, and the number of interior knots is given by $m_n - 4$. They propose a general *group SCAD* (gSCAD) procedure for selecting the groups of variables in a linear regression setting. Specifically, to select non-zero $\beta_k(t)$, they minimize the following PLS loss function

$$L(\beta, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{T} \left\{ y_{it} - \mu(t_j) - \sum_{k=1}^{p} \sum_{l=1}^{m_n} \beta_{kl} [B_l(t_j) X_{ik}] \right\}^2 + nT \sum_{k=1}^{p} p_\lambda(\|\boldsymbol{\beta}_k\|)$$

where $y_{it} = Y_i(t_j)$, $\mu = (\mu(t_1), ..., \mu(t_T))'$, $p_\lambda(\cdot)$ is the SCAD penalty with λ as a tuning parameter, and $\boldsymbol{\beta}_k = (\beta_{k1}, ..., \beta_{km_n})'$. An iterative algorithm based on local quadratic approximation of the non-convex penalty $p_\lambda(\|\boldsymbol{\beta}_k\|)$ as in Fan and Li (2001) is proposed. Under some overly restrictive conditions such as the knot locations are held *fixed* as the sample size increases, they generalize the arguments in Fan and Li (2001) to the group selection settings and establish the oracle property of gSCAD group selection procedure.

Wang, Li, and Huang (2008) consider a model similar to (5.4) but allow for unbalanced data:

$$Y_i(t_{ij}) = \sum_{k=1}^{p} \beta_k(t_{ij}) X_{ik}(t_{ij}) + \varepsilon_i(t_{ij}) \tag{5.5}$$

where $i = 1, ..., n$, $j = 1, ..., n_i$, $X_{ik}(t)$ is the covariate with time-varying effects, and the number of covariates p is fixed. They propose a PLS estimator using the SCAD penalty and basis expansion. The coefficients $\beta_k(t)$ can be approximated by a basis expansion $\beta_k(t) \approx \sum_{l=1}^{m_{nk}} \beta_{kl} B_{kl}(t)$ where various basis systems including Fourier bases, polynomial bases, and B-spline bases, can be used in the basis expansion to obtain $B_{kl}(t)$ for $l = 1, ..., m_{nk}$. Their objective function is given by

$$\sum_{i=1}^{n} w_i \sum_{j=1}^{T_i} \left\{ Y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{l=1}^{m_{nk}} \beta_{kl} [B_{kl}(t_j) X_{ik}(t_{ij})] \right\}^2 + \sum_{k=1}^{p} p_\lambda\left(\|\boldsymbol{\beta}_k\|_{\mathbf{R}_k}\right)$$

where the w_i's are weights taking value 1 if we treat all observations equally or $1/T_i$ if we treat each individual subject equally, $\|\boldsymbol{\beta}_k\|^2_{\mathbf{R}_k} = \boldsymbol{\beta}_k^{'} \mathbf{R}_k \boldsymbol{\beta}_k$, $\mathbf{R}_k$ is an $m_{nk} \times m_{nk}$ kernel matrix whose (i,j)-th element is given by

$$r_{k,ij} = \int_0^1 B_{ki}(t) B_{kj}(t) \, dt.$$

Under suitable conditions, they establish the theoretical properties of their procedure, including consistency in variable selection and the oracle property in estimation.

More recently, Wei, Huang, and Li (2011) also consider the model in (5.5) but allow the number of variables $p(= p_n)$ to be larger than n. They apply the group Lasso and basis expansion to simultaneously select the important variables and estimate the coefficient functions. The objective function of the group Lasso is given by

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{T_i} \left\{ Y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{l=1}^{m_{nk}} \beta_{kl} [B_{kl}(t_j) X_{ik}(t_{ij})] \right\}^2 + \sum_{k=1}^{p} \lambda \|\boldsymbol{\beta}_k\|_{\mathbf{R}_k}$$

where $\|\boldsymbol{\beta}_k\|_{\mathbf{R}_k}$ is defined as above. Under some conditions, they establish the estimation consistency of group Lasso and the selection consistency of adaptive group Lasso.

6 Variable Selection in Single Index Models

As a natural extension of linear regression model, the single index model (SIM) provides a specification that is more flexible than parametric models while retaining the desired properties of parametric models. It also avoids the curse of dimensionality through the index structure. Many methods have been proposed to estimate the coefficients in SIMs. Most of them can be classified into three categories. The first category includes the average derivative estimation method (Härdle and Stoker, 1989), the structure adaptive method (Hristache et al., 2001) and the outer product of gradients method (Xia, et al. 2002), which only focus on the estimation of unknown coefficients. The second category consists of methods that estimate the unknown link function and coefficients simultaneously, including Ichimura's (1993) semiparametric least square estimation and the minimum average conditional variance estimation (MAVE) by Xia et al. (2002). The third one is related to the inverse regression and is developed for sufficient dimension reduction (SDR), see, e.g., Li (1991).

Variable selection is a crucial problem in SIMs. Many classical variable selection procedures for linear regressions have been extended to SIMs. See, for example, Naik and Tsai (2001) and Kong and Xia (2008) for AIC and cross-validation. Based on the comparison of all the subsets of predictor variables, these methods are computationally intensive. Recently, Peng and Huang (2011) use penalized least squares method to estimate the model and select the significant variables simultaneously. Zeng, He and Zhu (2011) consider a Lasso-type approach called sim-Lasso for estimation and variable selection. Liang, Liu, Li and Tsai (2010) consider variable selection in partial linear single-indexed models using the SCAD penalty. Yang (2012) consider variable selection for functional index coefficient models. Some variable selection procedures are also proposed for generalized SIMs, see Zhu and Zhu (2009), Zhu, Qian, and Lin (2011) and Wang, Xu, and Zhu (2012).

6.1 Peng and Huang's (2011) penalized least squares for SIM

Peng and Huang (2011) consider the SIM

$$Y_i = g\left(X_i'\beta\right) + \varepsilon_i, \quad i = 1, 2, ..., n \tag{6.1}$$

where $g(\cdot)$ is a smooth unknown function, X_i is a $p \times 1$ vector of covariates, $\beta = (\beta_1, ..., \beta_p)'$ is a $p \times 1$ vector of parameters, and ε_i is a white noise with unknown variance σ^2. For identification, let $\|\beta\| = 1$ and $\text{sign}(\beta_1) = 1$ where $\text{sign}(a) = 1$ if $a > 0$, $= -1$ otherwise. They follow the idea of Carroll et al. (1997) and use an iterative algorithm to estimate β and the link function g simultaneously.

The unknown function $g(\cdot)$ can be approximated locally by a linear function

$$g(v) \approx g(u) + g'(u)(v - u)$$

when v lies in the neighborhood of u, and $g'(u) = dg(u)/du$. Given β, one can estimate $g(u)$ and $g'(u)$ by choosing (a, b) to minimize

$$\sum_{i=1}^{n} [Y_i - a - b(X_i'\beta - u)]^2 k_h(X_i'\beta - u) \tag{6.2}$$

where $k_h(\cdot) = k(\cdot/h)/h$ and $k(\cdot)$ is a symmetric kernel function. Let $\hat{g}(\cdot, \beta)$ denote the estimate of $g(u)$ given β. Given $\hat{g}(\cdot, \beta)$, one can estimate β by minimizing the following PLS function

$$\sum_{i=1}^{n} [Y_i - \hat{g}(X_i'\beta, \beta)]^2 + n\sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{6.3}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function with a tuning parameter λ. To solve the above nonlinear optimization problem, Peng and Huang propose to use the local approximation idea and update the estimate of β given the current estimates $\beta^{(0)}$ and $\hat{g}$ by minimizing the following penalized least squares function

$$\sum_{i=1}^{n} \left[Y_i - \hat{g}\left(X_i'\hat{\beta}^{(0)}, \hat{\beta}^{(0)}\right) - \hat{g}'\left(X_i'\hat{\beta}^{(0)}, \hat{\beta}^{(0)}\right) X_i'\left(\beta - \hat{\beta}^{(0)}\right)\right]^2 + n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{6.4}$$

The estimation procedure for β and $g(\cdot)$ is summarized as follows:

1. Obtain an initial estimate of β, say $\hat{\beta}^{(0)}$, by the least squares regression of Y_i on X_i. Let $\hat{\beta} = \hat{\beta}^{(0)}/\left\|\hat{\beta}^{(0)}\right\| \cdot \text{sign}(\hat{\beta}_1^{(0)})$ where $\hat{\beta}_1^{(0)}$ is the first element of $\hat{\beta}^{(0)}$.
2. Given $\hat{\beta}$, find $\hat{g}(u, \hat{\beta}) = \hat{a}$ and $\hat{g}'(u, \hat{\beta}) = \hat{b}$ by minimizing (6.2).
3. Update the estimate of β by minimizing (6.4) with $\hat{\beta}^{(0)}$ being replaced by $\hat{\beta}$.
4. Continue Steps 2-3 until convergence.
5. Given the final estimate $\hat{\beta}$ from Step 4, refine the estimate $\hat{g}(u, \hat{\beta})$ of $g(\cdot)$ by minimizing (6.2).

Peng and Huang argue that the above iterative algorithm can be regarded as an EM algorithm and different bandwidth sequence should be used in Steps 2-3 and 5. In Steps 2-3, one should assure the accuracy of the estimate of β and thus an undersmoothing bandwidth should be used to obtain the estimate of g; in step 5, one can obtain the final estimate of g by using the optimal bandwidth as if β were known. They discuss the choice of these two bandwidths and the tuning parameter λ as well. Under some conditions, they derive the convergence rate for $\hat{\beta}$ and show its oracle property.

6.2 Zeng, He, and Zhu's (2011) Lasso-type approach for SIMs

Zeng, He, and Zhu (2011) consider a Lasso-type approach called sim-Lasso for estimation and variable selection for the SIM in (6.2). The sim-Lasso method penalizes the derivative of the link function

and thus can be considered as an extension of the usual Lasso. They propose the following PLS minimization problem:

$$\min_{a,b,\beta,\|\beta\|=1} \sum_{j=1}^{n}\sum_{i=1}^{n}\left[Y_i - a_j - b_j\beta'\left(X_i - X_j\right)\right]^2 w_{ij} + \lambda\sum_{j=1}^{n}|b_j|\sum_{k=1}^{p}|\beta_k| \tag{6.5}$$

where $a = (a_1, ..., a_n)'$, $b = (b_1, ..., b_n)'$, $w_{ij} = K_h\left(X_i - X_j\right)/\sum_{l=1}^{n} K_h\left(X_l - X_j\right)$, $K_h(\cdot) = K(\cdot/h)/h^p$, $K(\cdot)$ is a kernel function and h is the bandwidth, and λ is the tuning parameter. Denote the objective function in (6.5) as $LM_\lambda(a, b, \beta)$ and its minimizer as $\hat{a}(\lambda) = (\hat{a}_1(\lambda), ..., \hat{a}_n(\lambda))'$, $\hat{b}(\lambda) = (\hat{b}_1(\lambda), ..., \hat{b}_n(\lambda))'$, and $\hat{\beta}(\lambda)$.

Note that the first part of $LM_\lambda(a, b, \beta)$ is the objective function for the MAVE estimation of β. Its inner summation is

$$\sum_{i=1}^{n}\left[Y_i - a_j - b_j\beta'\left(X_i - X_j\right)\right]^2 w_{ij}, \tag{6.6}$$

which is the least squares loss function for the local smoothing of g at $\beta' X_j$. A natural way to penalize (6.6) is to penalize the vector of linear coefficient $b_j\beta$ via the lasso type of penalty, yielding

$$\sum_{i=1}^{n}\left[Y_i - a_j - b_j\beta'\left(X_i - X_j\right)\right]^2 w_{ij} + \lambda|b_j|\sum_{k=1}^{p}|\beta_k|. \tag{6.7}$$

Summing (6.7) over i leads to the objective function in (6.5). The penalty term $\lambda\sum_{j=1}^{n}|b_j|\sum_{k=1}^{p}|\beta_k|$ has two-fold impact on the estimation of β. First, as the usual Lasso, it make $\hat{\beta}(\lambda)$ sparse and thus performs variable selection. Second, it also enforces shrinkage in $\hat{b}(\lambda)$ and may shrink some $\hat{b}_i(\lambda)$'s to zero. The second point is important as when g is relatively flat, its derivative is close to zero and does not contain much information about β.

Given β and b, the target function $LM_\lambda(a, b, \beta)$ can be minimized by $a_j = \tilde{Y}_i - b_j\beta'\left(\tilde{X}_i - X_j\right)$, where $\tilde{Y}_i = \sum_{i=1}^{n} w_{ij}Y_j$ and $\tilde{X}_i = \sum_{i=1}^{n} w_{ij}X_j$. Then $LM_\lambda(a, b, \beta)$ can be simplified to

$$\begin{aligned} L_\lambda(b, \beta) &= \min_a LM_\lambda(a, b, \beta) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{n}\left[Y_i - \tilde{Y}_i - b_j\beta'\left(X_i - \tilde{X}_j\right)\right]^2 w_{ij} + \lambda\sum_{j=1}^{n}|b_j|\sum_{k=1}^{p}|\beta_k|. \end{aligned}$$

When β is fixed, the target function L_λ is decoupled into n separate target functions, that is, $L_\lambda = \sum_{i=1}^{n} L_{\lambda,i}$, where

$$L_{\lambda,i} = \sum_{i=1}^{n}\left[Y_i - \tilde{Y}_i - b_j\beta'\left(X_i - \tilde{X}_j\right)\right]^2 w_{ij} + \lambda^*_\beta|b_j|$$

and $\lambda^*_\beta = \lambda\sum_{k=1}^{p}|\beta_k|$. The solution is

$$\hat{b}_j = \text{sgn}(\beta' R_j)\left(\frac{|\beta' R_j| - \lambda\sum_{k=1}^{p}|\beta_k|/2}{\beta' S_j\beta}\right)^+ \tag{6.8}$$

where

$$R_j = \sum_{i=1}^{n}\left(Y_i - \tilde{Y}_j\right)\left(X_i - \tilde{X}_j\right)w_{ij} \text{ and } S_j = \sum_{i=1}^{n}\left(X_i - \tilde{X}_j\right)\left(X_i - \tilde{X}_j\right)' w_{ij}.$$

For fixed b, minimizing $L_\lambda(b,\beta)$ becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{j=1}^{n}\sum_{i=1}^{n}\left[Y_i - \tilde{Y}_i - b_j\beta'\left(X_i - \tilde{X}_j\right)\right]^2 w_{ij} + \lambda_b^* \sum_{k=1}^{p}|\beta_k| \tag{6.9}$$

where $\lambda_b^* = \lambda\sum_{j=1}^{n}|b_j|$. It can be solved by the LARS-Lasso algorithm. The algorithm is summarized as follows:

1. Get an initial estimate $\hat{\beta}$ of β.
2. Given $\hat{\beta}$, calculate $\hat{b}_j$ as (6.8).
3. Given $\hat{b} = (\hat{b}_1, ..., \hat{b}_n)'$, use the LARS-Lasso algorithm to solve (6.9).
4. Renormalize $\hat{b}$ to $||\hat{\beta}||\hat{b}$ and $\hat{\beta}$ to $\hat{\beta}/||\hat{\beta}||$ and use them as $\hat{b}$ and $\hat{\beta}$ below.
5. Repeat Steps 2-4 until $(\hat{\beta}, \hat{b})$ converges.

Zeng, He and Zhu propose to use 10-fold cross-validation procedure to choose the penalty parameter λ and use the rule of thumb for bandwidth. They focus on the computational aspect of sim-Lasso but have not established its theoretical properties. They conjecture that by choosing the penalty parameter λ properly the sim-lasso possesses the usual consistency and convergence rate, but admit that the proof is nontrivial due to the interaction between the bandwidth h and the penalty parameter λ.

6.3 Liang, Liu, Li, and Tsai's (2010) partially linear single-index models

Liang, Liu, Li, and Tsai (2010) consider the following partially linear single-index model (PLSIM)

$$Y_i = \eta(Z_i'\boldsymbol{\alpha}) + X_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, ..., n \tag{6.10}$$

where Z_i and X_i are q-dimensional and p-dimensional covariate vectors, respectively, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)'$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, $\eta(\cdot)$ is an unknown differentiable function, ε_i is random error with zero mean and finite variance σ^2, and (X_i, Z_i) and ε_i are independent. They assume that $\|\boldsymbol{\alpha}\| = 1$ and α_1 is positive for identification. They propose a profile least-squares procedure to estimate the model and the SCAD penalty to select the significant variables.

Let $Y_i^* = Y_i - X_i'\boldsymbol{\beta}$ and $\Lambda_i = Z_i'\boldsymbol{\alpha}$. For given $\boldsymbol{\xi} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, $\eta(\cdot)$ can be estimated by the local linear regression to minimize

$$\sum_{i=1}^{n}\left[Y_i - a - b(\Lambda_i - u) - X_i'\boldsymbol{\beta}\right]^2 k_h(\Lambda_i - u),$$

with respect to a and b, where k_h is defined as before. Let $(\hat{a}, \hat{b})$ denote the minimizer. Then the profile estimator of $\eta(\cdot)$ is given by

$$\hat{\eta}(u;\boldsymbol{\xi}) = \hat{a} = \frac{K_{20}(u,\boldsymbol{\xi})K_{01}(u,\boldsymbol{\xi}) - K_{10}(u,\boldsymbol{\xi})K_{11}(u,\boldsymbol{\xi})}{K_{00}(u,\boldsymbol{\xi})K_{20}(u,\boldsymbol{\xi}) - K_{10}^2(u,\boldsymbol{\xi})}$$

where $K_{lj}(u,\boldsymbol{\xi}) = \sum_{i=1}^{n} k_h(\Lambda_i - u)(\Lambda_i - u)^l (X_i'\boldsymbol{\beta} - Y_i)^j$ for $l = 0,1,2$ and $j = 0,1$.They consider a PLS function

$$L_p(\boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^{n}[Y_i - \hat{\eta}(Z_i'\boldsymbol{\alpha};\boldsymbol{\xi}) - X_i'\boldsymbol{\beta}]^2 + n\sum_{j=1}^{q} p_{\lambda_{1j}}(|\alpha_j|) + n\sum_{k=1}^{p} p_{\lambda_{2k}}(|\beta_k|)$$

where $p_\lambda(\cdot)$ is a penalty function with a regularization parameter λ. Different penalty functions for different elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are allowed. If one wants to select variables in X_i only, one can set $p_{\lambda_{2k}}(\cdot) = 0$ for $k = 1,...,p$. Similarly, if one wants to select variables in Z_i only, one can set $p_{\lambda_{1j}}(\cdot) = 0$ for $j = 1,...,q$.

Because it is computationally expensive to minimize a criterion function with respect to $(p+q)$-dimensional regularization parameters, Liang, Liu, Li, and Tsai follow the approach of Fan and Li (2004) and set $\lambda_{1j} = \lambda SE(\hat{\alpha}_j^u)$ and $\lambda_{2k} = \lambda SE(\hat{\beta}_k^u)$, where λ is the tuning parameter, and $SE(\hat{\alpha}_j^u)$ and $SE(\hat{\beta}_k^u)$ are the standard errors of the unpenalized profile least squares estimators of α_j and β_k, respectively, for $j = 1,...,q$ and $k = 1,...,p$. Then they propose to use the SCAD penalty and select λ by minimizing the BIC-like criterion function given by

$$BIC(\lambda) = \log\{MSE(\lambda)\} + \frac{\log n}{n} df_\lambda$$

where $MSE(\lambda) = n^{-1}\sum_{i=1}^{n}\left[Y_i - \hat{\eta}(Z_i'\hat{\boldsymbol{\alpha}}_\lambda;\hat{\boldsymbol{\xi}}_\lambda) - X_i'\hat{\boldsymbol{\beta}}_\lambda\right]^2$, $\hat{\boldsymbol{\xi}}_\lambda = (\hat{\boldsymbol{\alpha}}_\lambda', \hat{\boldsymbol{\beta}}_\lambda')'$ is the SCAD estimator of $\boldsymbol{\xi}$ by using the tuning parameter λ, and df_λ is the number of nonzero coefficients in $\hat{\boldsymbol{\xi}}_\lambda$. They show that the BIC tuning parameter selector enables us to select the true model consistently and their estimate enjoys the oracle properties under some mild conditions.

In addition, it is worth mentioning that Liang and Wang (2005) consider the PLSIM in (6.10) when X_i is measured with additive error: $W_i = X_i + U_i$, where U_i is independent of (Y_i, X_i, Z_i). They propose two kernel estimation methods for this model but do not discuss the variable selection issue.

6.4 Yang's (2012) variable selection for functional index coefficient models

Yang (2012) considers the following functional index coefficient model (FICM) of Fan, Yao, and Cai (2003)

$$Y_i = g\left(\beta' Z_i\right)' X_i + \varepsilon_i \tag{6.11}$$

where $i = 1,...,n$, $X_i = (X_{i1},...,X_{ip})'$ is a $p \times 1$ vector of covariates, Z_i is a $q \times 1$ vector of covariate, ε_i is an error term with mean zero and variance σ^2, $\beta = (\beta_1,...,\beta_q)'$ is a $q \times 1$ vector of unknown parameters, and $g(\cdot) = (g_1(\cdot),...,g_p(\cdot))'$ is a vector of p-dimensional unknown functional coefficients. Assume that $\|\beta\| = 1$ and the first element β_1 of β is positive for identification. The sparsity of the model may come from two aspects: some of the functional index coefficients, $g_j(\cdot)$, $j = 1,...,p$, are identically zero; and some elements of β are zero. Yang proposes a two-step approach to select the significant covariates with functional coefficients, and then variable selection is applied to choose local significant variables with parametric coefficients. The procedure goes as follows:

1. Given a $\sqrt{n}$-consistent initial estimator $\hat{\beta}^{(0)}$ of β (e.g., that of Fan, Yao and Cai (2003)), we minimize the penalized local least squares to obtain the estimator $\hat{g}$ of g : $\hat{g} = \arg\min_{g} Q_h(g, \hat{\beta}^{(0)})$, where

$$Q_h(g,\beta) = \sum_{j=1}^{n}\sum_{i=1}^{n}\left[Y_i - g\left(\beta' Z_i\right)' X_i\right]^2 k_h\left(\beta' Z_i - \beta' Z_j\right) + n\sum_{l=1}^{p} p_{\lambda_l}\left(\|g_{l,\beta}\|\right)$$

$k_h(z) = k(z/h)/h$, k is a kernel function, h is a bandwidth parameter, $g_{l,\beta} = \left(g_l\left(\beta' Z_1\right), ..., g_l\left(\beta' Z_n\right)\right)'$ and $p_{\lambda_l}(\cdot)$ is the SCAD penalty function with tuning parameter λ_l.

2. Given the estimator $\hat{g}$ of g, we minimize the penalized global least squares objective function $Q(\beta, \hat{g})$ to obtain an updated estimator $\hat{\beta}$ of β, where

$$Q(\beta,\hat{g}) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \hat{g}\left(\beta' Z_i\right)' X_i\right]^2 + n\sum_{k=1}^{q} p_{\lambda_n}\left(|\beta_k|\right).$$

and $p_{\lambda_n}(\cdot)$ is the SCAD penalty function with tuning parameter λ_n.

Note that if one uses the Lasso penalty ($p_{\lambda_k}(a) = \lambda_k |a|$), the objective function in the first step becomes the penalized least squares criterion function used by Wang and Xia (2009). Yang proposes to choose the tuning parameters to minimize a BIC-type criterion function. Assuming that both p and q are fixed, he studies the consistency, sparsity, and the oracle property of the resulting functional index coefficient estimators $\hat{g}(\hat{\beta}' z)$ and $\hat{\beta}$. He applies his methodology to both financial and engineering data sets.

6.5 Generalized single index models

Zhu and Zhu (2009) consider estimating the direction of β and selecting important variables simultaneously in the following generalized single index model (GSIM) proposed by Li and Duan (1989) and Li (1991):

$$Y_i = G\left(X_i'\beta, \varepsilon_i\right) \tag{6.12}$$

where $G(\cdot)$ is an unknown link function, X_i is a $p \times 1$ vector of covariates, and ε_i is an error term that is independent of X_i. They allow $p = p_n$ to diverge as the sample size $n \to \infty$. The model in (6.12) is very general and covers the usual SIM and the heteroskedastic SIM (e.g., $Y = g_1(X'\beta) + g_2(X'\beta)\varepsilon$) as two special cases. Assume that $E(X_i) = 0$ and let $\Sigma = \text{Cov}(X_i)$.

Let $F(y)$ denote the cumulative distribution function (CDF) of the continuous response variable Y_i. Define

$$\beta^* = \arg\min_{b} E\left[l\left(b'X, F(Y)\right)\right]$$

where $l(b'X, F(Y)) = -F(Y)\, b'X + \psi(b'X)$ is a loss function and $\psi(\cdot)$ is a convex function. They show that under the sufficient recovery condition (which intuitively requires $E\left(X|\beta'X\right)$ to be linear in $\beta'X$: : $E\left(X|\beta'X\right) = \Sigma\beta\left(\beta'\Sigma\beta\right)^{-1}\beta'X)$ β^* identifies β in model (6.12) up to a multiplicative scalar. The main requirement for such an identification is that $E\left[l\left(b'X, F(Y)\right)\right]$ has a proper minimizer. This

condition relates to the unknown link function $G(\cdot)$ and is typically regarded as mild and thus widely assumed in the literature on SDR. To exclude the irrelevant regressors in the regression, Zhu and Zhu (2009) propose to estimate β as follows

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} l\left(b' X_i, F_n\left(Y_i\right)\right) + n \sum_{j=1}^{p} p_{\lambda_n}\left(|b_j|\right)$$

where $F_n(y) = n^{-1} \sum_{i=1}^{n} 1(y_i \leq y)$ is the empirical distribution function (EDF) of Y_i, b_j is the jth coordinate of b, and $p_{\lambda_n}(\cdot)$ is a penalty function with tuning parameter λ_n.

The loss function $l\left(b' X, F(Y)\right)$ covers the least squares measure as a special case, that is, $l\left(b' X_i, F\left(Y_i\right)\right)$ $= \left[b' X_i - F\left(Y_i\right)\right]^2 / 2$ by letting $\psi\left(b' X_i\right) = \left[b' X_i X_i' b + F^2\left(Y_i\right)\right] / 2$. Then the least square estimation is

$$\beta_{LS}^{*} = \underset{b}{\operatorname{argmin}} E\left[l\left(b' X_i, F\left(Y_i\right)\right)\right] = \underset{b}{\operatorname{argmin}} E\left[F\left(Y_i\right) - X_i' b\right]^2 = \Sigma^{-1} \operatorname{Cov}\left(X_i, F\left(Y_i\right)\right).$$

The sample version of the least squares estimate is given by

$$\hat{\beta}_{LS} = \underset{b}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \left[b' X_i - F_n\left(Y_i\right)\right]^2 / 2 + n \sum_{j=1}^{p} p_{\lambda_n}\left(|b_j|\right).$$

Zhu and Zhu (2009) suggest using the SCAD penalty. Under some regular conditions, they show that $\hat{\beta}_{LS}$ enjoys the oracle properties.

Zhu, Qian, and Lin (2011) follow the idea of Zhu and Zhu (2009) and propose a kernel-based method which automatically and simultaneously selects important predictors and estimates the direction of β in (6.12). As in Zhu and Zhu (2009), they also assume that $E(X_i) = 0$ and use Σ to denote $\operatorname{Cov}(X_i)$. The definition of the model in (6.12) is equivalent to saying that conditional on $\beta' X_i$, Y_i and X_i are independent. This implies the existence of a function $\tilde{f}$ such that

$$f\left(y|x\right) = \tilde{f}\left(y|\beta' x\right) \tag{6.13}$$

where $f\left(y|x\right)$ is the conditional probability density function (PDF) of Y_i given X_i and $\tilde{f}$ can be regarded as the conditional PDF of Y_i given $\beta' X_i$. (6.13), together with the chain rule, implies that

$$\frac{\partial f\left(y|x\right)}{\partial x} = \beta \frac{\partial \tilde{f}\left(y|\beta' x\right)}{\partial\left(\beta' x\right)}. \tag{6.14}$$

That is, the first derivative of $f\left(y|x\right)$ with respect to x is proportional to β. This motivates Zhu, Qian, and Lin (2011) to identify the direction of β through the derivative of the conditional PDF $f\left(y|x\right)$.

Let $k_h(u) = k(u/h)/h$ where $k(\cdot)$ is a univariate kernel function and h is a bandwidth parameter. Note that $f\left(y|x\right) \approx E\left[k_h\left(Y - y\right)|X = x\right] = E\left[k_h\left(Y - y\right)|\beta' X = \beta' x\right] \approx \tilde{f}\left(y|\beta' x\right)$ as $h \to 0$. When X is Gaussian with mean zero and covariance matrix Σ, a direct application of Stein's lemma yields that

$$H(y) = \Sigma^{-1} E\left[k_h\left(Y - y\right) X\right] \approx E\left[\frac{\partial f\left(y|X\right)}{\partial X}\right] \text{ as } h \to 0.$$

When the normality assumption does not hold, Zhu, Qian, and Lin (2011) relax it to the widely assumed linearity condition as in the sufficient recovery condition and show that $H(y)$ and thus $E\left[H(Y)\right]$ are

proportional to β for any fixed bandwidth h. Let

$$f^c(Y) = E\left[k_h\left(\tilde{Y} - Y\right)|Y\right] - E\left[k_h\left(\tilde{Y} - Y\right)\right] \tag{6.15}$$

where $\tilde{Y}$ is an independent copy of Y. They find that $E[H(Y)]$ is in spirit the solution to the following least squares minimization problem

$$\beta_0 = E[H(Y_i)] = \underset{b}{\text{argmin}} E\left[f^c(Y_i) - X_i'b\right]^2. \tag{6.16}$$

Note that the sample analogue of $f^c(Y)$ is given by

$$\hat{f}^c(Y) = \frac{1}{n}\sum_{i=1}^{n} k_h(Y_i - Y) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k_h(Y_i - Y_j).$$

Then one can obtain the unpenalized estimate of β_0 by

$$\hat{\beta}_0 = \underset{b}{\text{argmin}} \sum_{i=1}^{n}\left[\hat{f}^c(Y_i) - X_i'b\right]^2.$$

The adaptive Lasso estimate of β_0 is given by

$$\hat{\beta}_{0,ALasso} = \underset{b}{\text{argmin}} \sum_{i=1}^{n}\left[\hat{f}^c(Y_i) - X_i'b\right]^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |b_j|$$

where $\hat{w}_j = \left|\hat{\beta}_{0,j}\right|^{-\gamma}$ and $\hat{\beta}_{0,j}$ is the jth element of $\hat{\beta}_0$ for $j = 1, ..., p$. Assuming that p is fixed, Zhu, Qian, and Lin establish the oracle properties of $\hat{\beta}_{0,ALasso}$ under some regularity conditions.

Wang, Xu, and Zhu (2012) consider the variable selection and shrinkage estimation for several parametric and semiparametric models with the single-index structure by allowing $p = p_n$ to be divergent with n. Let $\delta = \text{Cov}(X_i, g(Y_i))$ for any function g. Define $\beta_g = \Sigma^{-1}\delta$. Under the assumption $E\left(X|\beta'X\right)$ is linear in $\beta'X$, Theorem 2.1 in Li (1991) immediately implies that β_g is proportional to β, i.e., $\beta_g = \kappa_g\beta$ for some constant κ_g. The least squares index estimate of β_g is given by

$$\hat{\beta}_g = \underset{b}{\text{argmin}} \sum_{i=1}^{n}\left[g(Y_i) - X_i'b\right]^2.$$

They propose a response-distribution transformation by replacing g by the CDF $F(y)$ of Y minus $1/2$. Since F is unknown in practice, they suggest using its EDF F_n and define the distribution-transformation least squares estimator as

$$\hat{\beta}_{F_n} = \underset{b}{\text{argmin}} \sum_{i=1}^{n}\left[F_n(Y_i) - \frac{1}{2} - X_i'b\right]^2.$$

The penalized version is given by

$$\hat{\beta}_{F_n} = \underset{b}{\text{argmin}} \sum_{i=1}^{n}\left[F_n(Y_i) - \frac{1}{2} - X_i'b\right]^2 + \sum_{j=1}^{p} p_{\lambda_n}\left(\left|\beta_j\right|\right),$$

where $p_{\lambda_n}(\cdot)$ can be the SCAD penalty or the MC penalty of Zhang (2010). They establish the selection consistency by allowing $p = p_n$ to grow at any polynomial rate under some moment conditions for X_i. If X_i's are normally distributed, it also allows p_n to grow exponentially fast.

7 Variable/Component Selection in General Nonparametric Models

In the previous four sections we review variable selection in semiparametric and nonparametric regression models that impose certain structures to alleviate the notorious "curse of dimensionality" problem in the literature. In this section we review variable selection in general nonparametric models that do not assume these structures. Even so, we remark that it is frequently assumed that certain decomposition of the general nonparametric regression functions exists, in which case the latter also exhibits a specific additive structure.

The literature on variable or component selection in general nonparametric models can be classified into two categories. The first category is carried out in the framework of *Smoothing Spline ANalysis Of VAriance* (SS-ANOVA) or *global* function approximation; see, e.g., Lin and Zhang (2006), Bunea (2008), Storlie, Bondell, Reich and Zhang (2011), and Comminges and Dalayan (2011). Lin and Zhang (2006) propose a new method called *COmponent Selection and Smoothing Operator* (COSSO) for model selection and model fitting in multivariate nonparametric regression models, in the framework of *smoothing spline ANOVA* (SS-ANOVA). As Huang, Breheny and Ma (2012) remark, the COSSO can be viewed as a group Lasso procedure in a reproducing kernel Hilbert space. Storlie, Bondell, Reich and Zhang (2011) propose the *adaptive COSSO* (ACOSSO) to improve the performance of COSSO. Bunea (2008) investigates the consistency of selection via the Lasso method in regression models, where the regression function is approximated by a given dictionary of M functions. Comminges and Dalayan's (2011) consider consistent variable selection in high dimensional nonparametric regression based on an orthogonal Fouries expansion of the regression function. The second category focuses on *local* selection of significant variables. Bertin and Lecué (2008) implement a two-step procedure to reduce the dimensionality of a local estimate. Lafferty and Wasserman (2008) introduce the Rodeo procedure, which attempts to assign adaptive bandwidths based on the derivative of kernel estimate with respect to the bandwidth for each dimension. Miller and Hall (2010) propose a method called LABAVS in local polynomial regression to select the variables and estimate the model.

7.1 Lin and Zhang's (2006) COSSO

Lin and Zhang (2006) consider the nonparametric regression

$$Y_i = f(X_i) + \varepsilon_i,\ i = 1, ..., n, \tag{7.1}$$

where f is the regression function to be estimated, $X_i = (X_{i1}, ..., X_{ip})' \in \mathcal{X} = [0,1]^p$ are p-dimensional vectors of covariates, and ε_i is independent noise with mean zero and finite variance σ^2. In the framework of SS-ANOVA, f exhibits the decomposition

$$f(x) = b + \sum_{j=1}^{p} f_j(x_j) + \sum_{1 \le j < k \le p} f_{jk}(x_j, x_k) + \cdots \tag{7.2}$$

where $x = (x_1, ..., x_p)'$, b is a constant, f_j's are the main effects, f_{jk} are the two-way interactions, and so on. The sequence is usually truncated somewhere to enhance interpretability. One can assure the

identifiability of the terms in (7.2) by some side conditions through averaging operators.

Let $\mathcal{F}$ be the *reproducing kernel Hilbert space* (RKHS) corresponding to the decomposition in (7.2). For the definition of RKHS, see Wahba (1991). Frequently $\mathcal{F}$ is a space of functions with a certain degree of smoothness, e.g., the second order Sobolev space, $\mathcal{S}^2 = \{g : g,\, g' \text{ are absolutely continuous and } g'' \in L^2[0,1]\}$. Let $\mathcal{H}_j$ be a function space of functions of x_j over $[0,1]$ such that $\mathcal{H}_j = \{1\} \oplus \bar{\mathcal{H}}_j$. Then $\mathcal{F}$ is the tensor product space of $\mathcal{H}_j$,

$$\mathcal{F} = \otimes_{j=1}^{p} \mathcal{H}_j = \{1\} \oplus \sum_{j=1}^{p} \bar{\mathcal{H}}_j \oplus \sum_{j<k} \left(\bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k\right) \oplus \cdots . \tag{7.3}$$

Each functional component in the SS-ANOVA decomposition (7.2) lies in a subspace in the orthogonal decomposition (7.3) of $\otimes_{j=1}^{p} \mathcal{H}_j$. But in practice the higher-order interactions are usually truncated for convenience to avoid the curse of dimensionality. In the simplest case where $f(x) = b + \sum_{j=1}^{p} f_j(x_j)$ with $f_j \in \bar{\mathcal{H}}_j$, the selection of functional components is equivalent to variable selection. In the general SS-ANOVA models model selection amounts to the selection of main effects and interaction terms in the SS-ANOVA decomposition. A general expression for the truncated space can be written as

$$\mathcal{F} = \{1\} \otimes \left\{\otimes_{j=1}^{q} \mathcal{F}^j\right\} \tag{7.4}$$

where $\mathcal{F}^1, ..., \mathcal{F}^q$ are q orthogonal subspaces of $\mathcal{F}$. $q = p$ gives the special case of additive models. When only main effects and two-way interaction effects are retained, the truncated space has $q = p(p+1)/2$, which includes p main effect spaces and $p(p-1)/2$ two-way interaction spaces.

Denote the norm in the RKHS $\mathcal{F}$ by $\|\cdot\|$. A traditional smoothing spline type method finds $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} [Y_i - f(X_i)]^2 + \lambda \sum_{j=1}^{q} \theta_j^{-1} \left\|P^j f\right\|^2 \tag{7.5}$$

where $P^j f$ is the orthogonal projection of f onto $\mathcal{F}^j$ and $\theta_j \geq 0$. If $\theta_j = 0$, the minimizer is taken to satisfy $\left\|P^j f\right\|^2 = 0$. The COSSO procedure finds $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} [Y_i - f(X_i)]^2 + \tau_n^2 J(f) \text{ with } J(f) = \sum_{j=1}^{q} \left\|P^j f\right\| \tag{7.6}$$

where τ_n is a smoothing parameter. The penalty term $J(f)$ is a sum of RKHS norms, instead of the squared RKHS norm penalty employed in the smoothing spline. $J(f)$ is a convex functional, which ensures the existence of the COSSO estimate. Let $\hat{f} = \hat{b} + \sum_{j=1}^{q} \hat{f}_j$ be a minimizer of (7.6).

Lin and Zhang (2006) form $\mathcal{F}$ using $\mathcal{S}^2$ with squared norm

$$\|g\|^2 = \left(\int_0^1 g(u)\,du\right)^2 + \left(\int_0^1 g'(u)\,du\right)^2 + \left(\int_0^1 g''(u)\,du\right)^2 \tag{7.7}$$

for each of the $\mathcal{H}_j$ in (7.3). They show that an equivalent expression of (7.6) is

$$\frac{1}{n}\sum_{i=1}^{n} [Y_i - f(X_i)]^2 + \lambda_0 \sum_{j=1}^{q} \theta_j^{-1} \left\|P^j f\right\|^2 + \lambda \sum_{j=1}^{q} \theta_j \text{ subject to } \theta_j \geq 0, \tag{7.8}$$

for $j = 1, ..., p$, where λ_0 is a constant and λ is a smoothing parameter. The constant λ_0 can be fixed at any positive value. For fixed θ, the COSSO (7.8) is equivalent to the smoothing spline (7.5). From the smoothing spline literature, it is well known that the solution f has the form

$$f(x) = \sum_{i=1}^{n} c_i R_\theta(X_i, x) + b$$

where $c = (c_1, ..., c_n)'$ and $R_\theta = \sum_{j=1}^{q} \theta_j R_j$ with R_j being the reproducing kernel of $\mathcal{F}_j$ (the $n \times n$ matrix $\{R_j(X_i, X_k)\}_{i,k=1}^{n}$). Then $f(x) = R_\theta c + b\mathbf{1}_n$, where $\mathbf{1}_n$ is an $n \times 1$ vector of ones. The problem (7.6) can be written as

$$\frac{1}{n}\left(Y - \sum_{j=1}^{q} \theta_j R_j c - b\mathbf{1}_n\right)'\left(Y - \sum_{j=1}^{q} \theta_j R_j c - b\mathbf{1}_n\right) + \lambda_0 \sum_{j=1}^{q} \theta_j c' R_j c + \lambda \sum_{j=1}^{q} \theta_j \tag{7.9}$$

where $Y = (Y_1, ..., Y_n)'$, and $\theta_j \geq 0$ for $j = 1, ..., q$. For fixed θ, (7.9) can be written as

$$\min_{c,b} (y - R_\theta c - b\mathbf{1}_n)'(y - R_\theta c - b\mathbf{1}_n) + n\lambda_0 c' R_\theta c$$

Then c and b can be solved as in Wahba (1991). On the other hand, if c and b are fixed, let $g_j = R_j c$ and G be the matrix with the jth column being g_j. θ that minimizes (7.9) is the solution to

$$\min_{\theta} (z - G\theta)'(z - G\theta) + n\lambda \sum_{j=1}^{q} \theta_j \text{ subject to } \theta_j \geq 0 \text{ for } j = 1, .., q$$

or

$$\min_{\theta} (z - G\theta)'(z - G\theta), \text{ subject to } \theta_j \geq 0, j = 1, ..., q, \text{ and } \sum_{j=1}^{q} \theta_j \leq M$$

where $z = y - (1/2) n\lambda_0 c - b\mathbf{1}_n$ and M is a positive constant. The tuning parameter can be chosen by 5-fold or 10-fold cross-validation. Lin and Zhang study the theoretical properties such as the existence and rate of convergence of the COSSO estimator.

In the framework of SS-ANOVA, Zhang and Lin (2006) study the component selection and smoothing for nonparametric regression in the more general setting of exponential family regression, and Leng and Zhang (2006) study the same issue for a nonparametric extension of the Cox proportional hazard model. The former allows the treatment of non-normal responses, binary and polychotomous responses, and event counts data. The latter demonstrates great flexibility and easy interpretability in modeling relative risk functions for censored data.

7.2 Storlie, Bondell, Reich, and Zhang's (2011) ACOSSO

The oracle properties used before are mainly defined for the finite dimensional parameter in parametric or semiparametric models. In the context of nonparametric regression, Storlie, Bondell, Reich, and Zhang (2011) extend this notion by saying a nonparametric regression estimator has the nonparametric *weak (np)-oracle property* if it selects the correct subset of predictors with probability tending to one,

and estimates the regression function at the optimal nonparametric rate. Note that the *strong* version of the oracle property requires the estimator should have the asymptotic distribution as the oracle one. The SS-ANOVA-based COSSO procedure has not been demonstrated to possess the weak np-oracle property. Instead, it has a tendency to over-smooth the nonzero functional components in order to set the unimportant functional components to zero. Storlie, Bondell, Reich, and Zhang propose the adaptive COSSO (ACOSSO) which possesses the weak np-oracle properties.

Like Lin and Zhang (2006), Storlie, Bondell, Reich, and Zhang consider the nonparametric regression model in (7.1) where $X_i = (X_{i1}, ..., X_{ip})' \in \mathcal{X} = [0,1]^p$, ε_i's are independent of X_i and are uniformly sub-Gaussian with zero mean. They obtain their estimate of the function $f \in \mathcal{F}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - f\left(x_i\right)\right]^2 + \lambda\sum_{j=1}^{p} w_j \left\|P^j f\right\| \tag{7.10}$$

where $0 < w_j \leq \infty$ are weights that can depend on an initial estimate of f, e.g., the COSSO estimate $\tilde{f}$. They suggest the choice

$$w_j = \left\|P^j \tilde{f}\right\|_{L_2}^{-\gamma} \text{ for } j = 1, ..., p,$$

where $\left\|P^j \tilde{f}\right\|_{L_2} = \{\int_{\mathcal{X}}[P^j \tilde{f}\left(x\right)]^2 dx\}^{1/2}$ and $\gamma > 0$. The tuning parameter is also chosen via 5-fold or 10-fold cross-validation. Under some regular conditions, they show their estimator possess the weak np-oracle property when $f \in \mathcal{F}$ is additive in the predictors so that $\mathcal{F} = \{1\} \oplus \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_p$ where each $\mathcal{F}_j$ is a space of functions corresponding to x_j.

7.3 Bunea's (2008) consistent selection via the Lasso

Bunea (2008) considers the approximation of the regression function f in (7.1) with elements of a given dictionary of M functions. Let

$$\Lambda = \left\{\lambda \in \mathbb{R}^M : \left\|f - \sum_{j=1}^{M}\lambda_j f_j\right\|^2 \leq C_f r_{n,M}^2\right\}$$

where $C_f > 0$ is a constant depending only on f and $r_{n,M}$ is a positive sequence that converges to zero. For any $\lambda = (\lambda_1, ..., \lambda_M)' \in \mathbb{R}^M$, let $J\left(\lambda\right)$ denote the index set corresponding to the non-zero components of λ and $M\left(\lambda\right)$ its cardinality. Let $p^* = \min\left\{M\left(\lambda\right) : \lambda \in \Lambda\right\}$. Define

$$\lambda^* = \underset{\lambda \in \mathbb{R}^M}{\arg\min}\left\{\left\|f - \sum_{j=1}^{M}\lambda_j f_j\right\|^2 : M\left(\lambda\right) = p^*\right\}.$$

Let $I^* = J\left(\lambda^*\right)$ denote the index set corresponding to the non-zero elements of λ^*. Note the cardinality of I^* is given by p^* and thus $f^* = \sum_{j \in I^*}\lambda_j^* f_j$ provides the sparest approximation to f that can be realized with $\lambda \in \Lambda$ and $\left\|f^* - f\right\|^2 \leq C_f r_{n,M}^2$. This motivates Bunea to treat I^* as the target index set.

Bunea considers estimating the set I^* via the l_1-penalized least squares. First, he computes

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^M} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \lambda_j f_j \left(X_i \right) \right]^2 + 2 \sum_{j=1}^{M} w_{nj} \left| \lambda_j \right| \right\}$$

where $w_{nj} = r_{n,M} \left\| f_j \right\|_n$ and $\left\| f_j \right\|_n^2 = n^{-1} \sum_{i=1}^{n} \left[f_j \left(X_i \right) \right]^2$. Let $\hat{I}$ denote the index set corresponding to the non-zero components of $\hat{\lambda}$. He shows that $P\left(\hat{I} = I^* \right) \to 1$ as $n \to \infty$ under some conditions in conjunction with the requirement that $p^* r_{n,M} \to 0$.

7.4 Comminges and Dalayan's (2011) consistent variable selection in high dimensional nonparametric regression

Comminges and Dalayan (2011) consider the general nonparametric regression model (7.1) where X_i's are assumed to take values in $[0,1]^p$, $E\left(\varepsilon_i | X_i \right) = 0$, $E\left(\varepsilon_i^2 | X_i \right) = \sigma^2$, and $p = p_n$ may diverge to the infinity with n. They assume that f is differentiable with a squared integrable gradient and that the density function $g\left(x \right)$ of X_i exists and is bounded away from 0 from below. Define the Fourier basis

$$\varphi_{\mathbf{k}} \left(x \right) = \begin{cases} 1 & \text{if } \mathbf{k} = \mathbf{0} \\ \sqrt{2} \cos \left(2\pi \mathbf{k}' x \right) & \text{if } \mathbf{k} \in \left(\mathbb{Z}^p \right)_+ \\ \sqrt{2} \sin \left(2\pi \mathbf{k}' x \right) & \text{if } -\mathbf{k} \in \left(\mathbb{Z}^p \right)_+ \end{cases}$$

where $\left(\mathbb{Z}^p \right)_+$ denotes the set of all $\mathbf{k} = \left(k_1, ..., k_p \right)' \in \mathbb{Z}^p \setminus \{0\}$ such that the first nonzero element of $\mathbf{k}$ is positive. Let

$$\Sigma_L = \left\{ f : \sum_{\mathbf{k} \in \mathbb{Z}^p} k_j \ < f, \varphi_{\mathbf{k}} >^2 \leq L \ \forall j \in \{1, ..., p\} \right\}$$

where $< \cdot, \cdot >$ stands for the scalar product in $L^2\left([0,1]^p ; \mathbb{R} \right)$, i.e., $<a, b> \int_{[0,1]^p} a\left(x \right) b\left(x \right) dx$ for any $a, b \in L^2\left([0,1]^p ; \mathbb{R} \right)$. Comminges and Dalayan assume that the regression function f belongs to Σ_L and for some $J \subset \{1, ..., p\}$ of cardinality $p^* \leq p$, $f\left(x \right) = \bar{f}\left(x_J \right)$ for some $\bar{f} : \mathbb{R}^{|J|} \to \mathbb{R}$, and it holds that

$$Q_j \left[f \right] \equiv \sum_{\mathbf{k}: k_j \neq 0} \theta_{\mathbf{k}} \left[f \right]^2 \geq \kappa, \ \forall j \in J,$$

where $\theta_{\mathbf{k}} \left[f \right] =< f, \varphi_{\mathbf{k}} >$. Clearly, J refers to the sparsity pattern of f and $Q_j \left[f \right] = 0$ if $j \notin J$.

The Fourier coefficients $\theta_{\mathbf{k}} \left[f \right]$ can be estimated by their empirical counterparts

$$\hat{\theta}_{\mathbf{k}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\varphi_{\mathbf{k}} \left(X_i \right)}{g\left(X_i \right)} Y_i, \ \mathbf{k} \in \mathbb{Z}^p.$$

Let $S_{m,l} = \{\mathbf{k} \in \mathbb{Z}^p : \left\| \mathbf{k} \right\|_2 \leq m, \ \left\| \mathbf{k} \right\|_0 \leq l\}$ and $N\left(p^*, \gamma \right) =$Card$\{\mathbf{k} \in \mathbb{Z}^{p^*} : \left\| \mathbf{k} \right\|_2^2 \leq \gamma p^*, k_1 \neq 0\}$ where $l \in \mathbb{N}$ and $\gamma > 0$. Note that if $j \notin J$ then $\theta_{\mathbf{k}} \left[f \right] = 0$ for every $\mathbf{k}$ such that $k_j \neq 0$, and if $j \in J$ then there exists $\mathbf{k} \in \mathbb{Z}^p$ with $k_j \neq 0$ such that $\left| \theta_{\mathbf{k}} \left[f \right] \right| > 0$. Comminges and Dalayan define their estimator of J by

$$\hat{J}_n \left(m, \lambda \right) = \left\{ j \in \{1, ..., p\} : \max_{\mathbf{k} \in S_{m,p^*}: k_j \neq 0} \left| \hat{\theta}_{\mathbf{k}} \right| > \lambda \right\}.$$

They show that $P\left(\hat{J}_n(m,\lambda) \neq J\right) \leq 3(6mp)^{-p^*}$ under some regularity conditions related to $N(p^*,\gamma)$ and (p,p^*,n). It is possible for p^* to be either fixed or tend to the infinity as $n \to \infty$. Unfortunately, Comminges and Dalayan deliberately avoid any discussion on the computational aspects of the variable selection and focus exclusively on the consistency of variable selection without paying any attention to the consistency of regression function estimation. Two problems have to be addressed in order to implement their procedure, namely, the estimate of the typically unknown density function g and the determination of p^*.

7.5 Bertin and Lecué's (2008)

Bertin and Lecué (2008) consider the nonparametric regression model (7.1) where ε_i's are i.i.d. Gaussian random variables with variance σ^2 and independent of X_i's, and f is the unknown regression function. Suppose the nonparametric regression function f satisfies a sparseness condition:

$$f(x) = \bar{f}(x_{\mathbf{R}}) \tag{7.11}$$

where $x_{\mathbf{R}} = (x_j : j \in \mathbf{R})$, $\mathbf{R} \subset \{1,...,p\}$ is a subset of p covariates, of size $p^* = |\mathbf{R}| < p$. Obviously, $x_{\mathbf{R}}$ denotes the set of relevant variables. They are interested in the pointwise estimation of f at a fixed point $x = (x_1,...,x_p)'$ and the construction of some estimate $\hat{f}_n$ having the smallest pointwise integrated quadratic risk $E[\hat{f}_n(x) - f(x)]^2$.

Assume f to be β-Holderian around x with $\beta > 0$, denoted by $f \in \Sigma(\beta, x)$. A function $f : \mathbb{R}^p \to \mathbb{R}$ is β-Holderian at point x with $\beta > 0$ if (i) f is l-times differentiable in x ($l = \lfloor \beta \rfloor$); and (ii) there exists $L > 0$ such that for any $t = (t_1,...,t_n) \in B_\infty(x,1)$ (the unit l_∞-ball of center x and radius 1), $|f(t) - P_l(f)(t,x)| \leq L\|t-x\|_1^\beta$, where $P_l(f)(\cdot,x)$ is the Taylor polynomial of order l associated with f at point x. Assume that there exists a subset of $J = \{i_1,...,i_{p^*}\} \subset \{1,...,p\}$ such that

$$f(x_1,...,x_p) = \bar{f}(x_{i_1},...,x_{i_{p^*}}).$$

That is, the "real" dimension of the model is not p but p^*. Bertin and Lecué's goal is twofold: (i) determine the set of indices $J = \{i_1,...,i_{p^*}\}$, and (ii) construct an estimator of $f(x)$ that converges at rate $n^{-2\beta/(2\beta+p^*)}$ which is the fastest convergence rate when $f \in \Sigma(\beta,x)$ and the above sparsity condition is satisfied.

To determine the set of indices, based on the principle of local linear regression, they consider the following set of vectors

$$\bar{\Theta}_1(\lambda) = \underset{\theta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \frac{1}{nh^p} \sum_{i=1}^n \left[Y_i - U\left(\frac{X_i - x}{h}\right)' \theta \right]^2 K\left(\frac{X_i - x}{h}\right) + 2\lambda \|\theta\|_1 \right\}$$

where $U(v) = (1, v_1,...,v_p)'$ for any $v = (v_1,...,v_p)'$, $\theta = (\theta_0,\theta_1,...,\theta_p)$, $\|\theta\|_1 = \sum_{j=0}^p |\theta_j|$, h is a bandwidth, and $K(\cdot)$ is symmetric kernel function. The l_1 penalty makes the solution vector $\bar{\Theta}(\lambda)$ sparse and then selects the variables locally. Another selection procedure, which is close to the previous

one but requires less assumption on the regression function, is given by

$$\bar{\Theta}_2(\lambda) = \underset{\theta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \frac{1}{nh^p}\sum_{i=1}^{n}\left[Y_i + f_{\max} + Ch - U\left(\frac{X_i - x}{h}\right)'\theta\right]^2 K\left(\frac{X_i - x}{h}\right) + 2\lambda\|\theta\|_1 \right\}$$

where $C > 0$ is a constant, $f_{\max} > 0$ and $|f(x)| \leq f_{\max}$. Here, the response variable Y_i is translated by $f_{\max} + Ch$.

Let $\hat{J}_1$ and $\hat{J}_2$ be the subset of indices selected by the above procedures for a given λ. Based on these sets of indices, Bertin and Lecué consider a local polynomial regression of degree $l = \lfloor\beta\rfloor$ by regressing Y_i on the selected variables. Under different conditions on function f, they show that $\hat{J}_1 = J$ or $\hat{J}_2 = J$ with a probability approaching 1, and for $\hat{J}_2$ the local polynomial estimate in the second step can achieve the fastest convergence rate under some conditions. The selection is proved to be consistent when $p^* = O(1)$ but p is allowed to be as large as $\log n$, up to a multiplicative constant.

7.6 Lafferty and Wasserman's (2008) Rodeo procedure

Lafferty and Wasserman (2008) presented a greedy method for simultaneously preforming local bandwidth selection and variable selection in the nonparametric regression model (7.1) where ε_i's are i.i.d. Gaussian random variables with zero mean and variance σ^2. Suppose the nonparametric regression function f satisfies a sparseness condition in (7.11) with $p^* = |\mathbf{R}| \ll p$. Without loss of generality, we assume that $x_{\mathbf{R}} = (x_1, ..., x_{p^*})'$ so that the last $p - p^*$ elements in x are irrelevant. Based on the idea that bandwidth and variable selection can be simultaneously performed by computing the infinitesimal change in a nonparametric estimator as a function of smoothing parameters, Lafferty and Wasserman propose the general framework for the *regularization of derivative expectation operator* (Rodeo).

The key idea is as follows. Fix a point x and let $\hat{f}(x)$ be an estimator of $f(x)$ based on a vector of smoothing parameters $h = (h_1, ..., h_p)'$. Let $F(h) = E[\hat{f}_h(x)]$. Assume that $x = X_i$ is one of the observed data points and $\hat{f}_0(x) = Y_i$. In this case, $f(x) = F(0) = E(Y_i)$. If $P = (h(t) : 0 \leq t \leq 1)$ is a smooth path through the set of smoothing parameters with $h(0) = 0$ and $h(1) = 1$ (or any other fixed large bandwidth). Then

$$\begin{aligned} f(x) &= F(0) = F(1) + F(0) - F(1) \\ &= F(1) - \int_0^1 \frac{dF(h(s))}{ds} ds \\ &= F(1) - \int_0^1 D(h(s))' \dot{h}(s)\, ds, \end{aligned}$$

where $D(h(s)) = \nabla F(h) = \left(\frac{\partial F}{\partial h_1}, ..., \frac{\partial F}{\partial h_p}\right)'$ and $\dot{h}(s) = \frac{dh(s)}{ds}$. Noting that an unbiased estimator of $F(1)$ is $\hat{f}_1(x)$, an unbiased estimator of $D(h)$ is

$$Z(h) = \left(\frac{\partial \hat{f}_h(x)}{\partial h_1}, ..., \frac{\partial \hat{f}_h(x)}{\partial h_p}\right)'.$$

The naive estimator

$$\hat{f}(x) = \hat{f}_1(x) - \int_0^1 Z(h(s))' \dot{h}(s)\, ds$$

is equal to $\hat{f}_0(x) = Y_i$, which is a poor estimator because of the large variance of $Z(h)$ for small h. Nevertheless, the sparsity assumption on f suggests that $D(h)$ is also sparse for some paths. Then using an estimator $\hat{D}(h)$ which uses the sparsity assumption yields the following estimate of $f(x)$

$$\tilde{f}(x) = \hat{f}_1(x) - \int_0^1 \hat{D}(h(s))' \dot{h}(s)\, ds.$$

The implementation of such an estimator requires us to find a path for which the derivative $D(h)$ is also sparse, and then take advantage of this sparseness when estimating $D(h)$ along that path.

A key observation is that if x_j is irrelevant in x, then changing the bandwidth h_j should cause only a small change in $\hat{f}_h(x)$. Conversely, if x_j is relevant in x, then changing h_j should cause a large change in $\hat{f}_h(x)$. Thus $Z_j(h) = \partial \hat{f}_h(x) / \partial h_j$ should discriminate between relevant and irrelevant covariates. Let $h_j \in \mathcal{H} = \{h_0, \beta h_0, \beta^2 h_0, ...\}$ for some $\beta \in (0,1)$. A greedy version of estimator of $D_j(h)$, the jth element of $D(h)$, would set $\hat{D}_j(h) = 0$ when $h_j < \hat{h}_j$, where $\hat{h}_j$ is the first h such that $|Z_j(h)| < \lambda_j(h)$ for some threshold λ_j where $h = a$ for a scalar a means $h = (a, ..., a)'$, a $p \times 1$ vector. That is

$$\hat{D}_j(h) = Z_j(h)\, \mathbf{1}(|Z_j(h)| > \lambda_j(h)).$$

This greedy version, coupled with the hard threshold estimator, yields $\tilde{f}(x) = \hat{f}_{\hat{h}}(x)$ where $\hat{h} = (\hat{h}_1, ..., \hat{h}_p)'$. This is a bandwidth selection procedure based on testing.

For local linear regression, Lafferty and Wasserman give explicit expressions for $Z(h)$. The local linear estimator of $f(x)$ by using kernel K and bandwidth $h = (h_1, ..., h_p)'$ is given by

$$\hat{f}_h(x) = \sum_{i=1}^n G(X_i, x, h)\, Y_i$$

where $G(u, x, h) = e_1' (X_x' W_x X_x)^{-1} \begin{pmatrix} 1 \\ u - x \end{pmatrix} K_h(u - x)$ is the effective kernel, $e_1 = (1, 0, ..., 0)'$, $K_h(u) = (h_1 ... h_p)^{-1} K(u_1/h_1, ..., u_p/h_p)$, X_x is an $n \times (p+1)$ matrix whose ith row is given by $(1, (X_i - x)')$, and W_x is a diagonal matrix with (i,i)-element $K_h(X_i - x)$. In this case,

$$Z_j(h) = \frac{\partial \hat{f}_h(x)}{\partial h_j} = \sum_{i=1}^n \frac{\partial G(X_i, x, h)}{\partial h_j} Y_i.$$

Lafferty and Wasserman derive the explicit expression for $\partial G(X_i, x, h) / \partial h_j$ and $Z_j(h)$. Let

$$s_j = \mathrm{Var}(Z_j(h) | X_1, ..., X_n) = \sigma^2 \sum_{i=1}^n \left(\frac{\partial G(X_i, x, h)}{\partial h_j} \right)^2.$$

They illustrate how to perform the Rodeo via the hard thresholding as follows:

1. Select a constant $\beta \in (0,1)$ and the initial bandwidth $h_0 = c / \log\log n$.

2. Initialize the bandwidths, and activate all covariates: (a) $h_j = h_0$ for $j = 1, ..., p$, and (b) $\mathcal{A} = \{1, 2, ..., p\}$.

3. While $\mathcal{A}$ is nonempty, for each $j \in \mathcal{A}$: (a) compute Z_j and s_j; (b) compute threshold value $\lambda_j = s_j\sqrt{2\log n}$; and (c) if $|Z_j| > \lambda_j$, reassign select βh_j to h_j; otherwise remove j from $\mathcal{A}$.

4. Output the bandwidth $h^* = (h_1, ..., h_p)$ and estimator $\tilde{f}(x) = \hat{f}_{h^*}(x)$.

Under some conditions, the Rodeo outputs bandwidths h^* that satisfies $P(h_j^* = h_0 \text{ for all } j > p^*) \to 1$ where recall X_{ij}'s are irrelevant variables for all $j > p^*$. In particular, the Rodeo selection is consistent when $p = O(\log n / \log\log n)$ and its estimator achieves the near optimal minimax rate of convergence while p^* does not increase with n. Lafferty and Wasserman explain how to estimate σ^2 used in the definition of s_j, and obtain other estimators of $D(h)$ based on the soft thresholding.

7.7 Miller and Hall's (2010) LABAVS in local polynomial regression

Miller and Hall (2010) propose a flexible and adaptive approach to local variable selection using local polynomial regression. The key technique is careful adjustment of the local regression bandwidths to allow for variable redundancy. They call their method as LABAVS, standing for "*locally adaptive bandwidth and variable selection*". The model is as given in (7.1). They consider the local polynomial estimation of f at a fixed point x. Let $H = \text{diag}\left(h_1^2, ..., h_p^2\right)$. Let $K(x) = \Pi_{j=1}^p k(x_j)$ be the p-dimensional rectangular kernel formed from a univariate kernel k with support on $[-1, 1]$ such as the tricubic kernel: $k(v) = (35/32)\left(1 - v^2\right)^3 \mathbf{1}(|v| \leq 1)$. Let $K_H(x) = |H|^{-1/2} K\left(H^{-1/2}x\right)$. We write $H(x)$ when H varies as a function of x. Asymmetric bandwidths are defined as having a lower and an upper diagonal bandwidth matrix, H^L and H^U, respectively, for a given estimation point x. The kernel weight of an observation X_i at an estimation point x with asymmetrical local bandwidth matrices $H^L(x)$ and $H^U(x)$ is given by

$$K_{H^L(x),H^U(x)}(X_i - x) = \prod_{j:X_{ij}<x_j} h_j^L(x)^{-1} k\left(\frac{X_{ij} - x_j}{h_j^L(x)}\right) \times \prod_{j:X_{ij}\geq x_j} h_j^U(x)^{-1} k\left(\frac{X_{ij} - x_j}{h_j^U(x)}\right),$$

which amounts to having possibly different window sizes above and below x in each direction.

Miller and Hall's LABAVS algorithm works as follows:

1. Find an initial bandwidth matrix $H = \text{diag}\left(h^2, ..., h^2\right)$.

2. For each point x of a representative grid in the data support, perform local variable selection to determine disjoint index sets $\hat{A}^+(x)$ and $\hat{A}^-(x)$ for variables that are considered relevant and redundant, respectively. Note that $\hat{A}^+(x) \cup \hat{A}^-(x) = \{1, ..., p\}$.

3. For any given x, derive new local bandwidth matrices $H^L(x)$ and $H^U(x)$ by extending the bandwidth in each dimension indexed in $\hat{A}^-(x)$. The resulting space given nonzero weight by the kernel $K_{H^L(x),H^U(x)}(u - x)$ is the rectangle of maximal area with all grid points x_0 inside the rectangle satisfying $\hat{A}^+(x_0) \subset \hat{A}^+(x)$ where $\hat{A}^+(x)$ is calculated explicitly as in Step 2, or taken as the set corresponding the closet grid point to x.

4. Shrink the bandwidth slightly for those variables in $\hat{A}^{+}(x)$ according to the amount that bandwidths have increased in the other variables.

5. Compute the local polynomial estimate at x, excluding variables in $\hat{A}^{-}(x)$ and using adjusted asymmetrical bandwidths $H^{L}(x)$ and $H^{U}(x)$. For example, in the local linear regression case, one chooses a and b to minimize

$$\sum_{i=1}^{n}\left[Y_i - a - b'\left(X_i - x\right)\right]^2 K_{H^{L}(x),H^{U}(x)}\left(X_i - x\right).$$

Steps 2 and 4 of the above algorithm are referred to as the variable selection step and variable shrinkage step, respectively. Miller and Hall suggest three possible ways to select variables at x in Step 2, namely, hard thresholding, backwards stepwise approach, and local Lasso. Let

$$\bar{X}_{j,x} = \frac{\sum_{i=1}^{n} X_{ij} K_{H(x)}\left(X_i - x\right)}{\sum_{i=1}^{n} K_{H(x)}\left(X_i - x\right)}, \text{ and } \bar{Y}_x = \frac{\sum_{i=1}^{n} Y_i K_{H(x)}\left(X_i - x\right)}{\sum_{i=1}^{n} K_{H(x)}\left(X_i - x\right)},$$

which are the local standardization of the data at point x. Let $\tilde{Y}_i = \left(Y_i - \bar{Y}_x\right)\left[K_{H(x)}\left(X_i - x\right)\right]^{1/2}$, and $\tilde{X}_{ij} = \frac{\left(X_{ij} - \bar{X}_{j,x}\right)\left[K_{H(x)}(X_i - x)\right]^{1/2}}{\left[\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_{j,x}\right)^2 K_{H(x)}(X_i - x)\right]^{1/2}}$. In the local linear regression case, the hard thresholding method chooses parameters to minimize the weighted least squares

$$\sum_{i=1}^{n}\left[\tilde{Y}_i - \beta_0 - \sum_{j=1}^{p}\beta_j \tilde{X}_{ij}\right]^2,$$

and classifies as redundant variables for which $\left|\hat{\beta}_j\right| < \lambda$ for some tuning parameter λ, where $(\hat{\beta}_0, ..., \hat{\beta}_p)$ is the solution to the above minimization problem. The variable shrinkage step and Step 3 are fairly complicated and computationally demanding. We refer the readers directly to Miller and Hall (2010) who also compare their approach with other local variable selection approaches in Bertin and Lecué (2008) and Lafferty and Wasserman (2008). They establish the strong oracle property for their estimator.

8 Variable Selection in semiparametric/nonparametric Quantile Regression

As a generalization of least absolute deviation regression (LADR), quantile regression (QR) has attracted huge interest in the literature and has been widely used in economics and finance; see Koenker (2005) for an overview. To select the significant variables is an important problem for QR. Many procedures have been proposed. Koenker (2004) applies the Lasso penalty to mixed-effects linear QR model for longitudinal data to shrink the estimator of random effects. Wang, Li and Jiang (2007) consider linear LADR with the adaptive Lasso penalty. Zou and Yuan (2008) propose a model selection procedure based on composite linear QRs. Wu and Liu (2009) consider the SCAD and adaptive Lasso

in linear QR models. Belloni and Chernozhukov (2011a) consider l_1-penalized or post-l_1-penalized QR in high-dimensional linear sparse models. Liang and Li (2009) propose penalized QR (PQR) for PLMs with measurement error by using orthogonal regression to correct the bias in the loss function due to measurement error. Koenker (2011) considers the additive models for QR which include both parametric and nonparametric components. Kai, Li, and Zou (2011) consider efficient estimation and variable selection for semiparametric varying-coefficient PLM using composite QR. Lin, Zhang, Bondell, and Zou (2012) consider variable selection for nonparametric QR via SS-ANOVA. In this section, we focus on reviewing variable selection in semiparametric/nonparametric QR models.

8.1 Liang and Li's (2009) penalized quantile regression for PLMs with measurement error

Liang and Li (2009) consider the PLM in (4.1) when X_i is measured with additive error:

$$W_i = X_i + U_i \tag{8.1}$$

where U_i is the measurement error with mean zero and unknown covariance Σ_{uu} and U_i is independent of (X_i, Z_i, Y_i). They propose a penalized quantile regression (PQR) based on the orthogonal regression. That is, the objective function is defined as the sum of squares of the orthogonal distances from the data points to the straight line of regression function, instead of residuals from the classical regression. He and Liang (2000) apply the idea of orthogonal regression for QR for both linear and partially linear models with measurement error, but do not consider the variable selection problem. Liang and Li further use the orthogonal regression method to develop a PQR procedure to select significant variables in the PLMs.

To define orthogonal regression for QR with measurement error, it is assumed that the random vector $(\varepsilon_i, U_i')'$ follows an elliptical distribution with mean zero and covariance matrix $\sigma^2\Sigma$ where σ^2 is unknown, and Σ is a block diagonal matrix with $(1,1)$-element being 1 and the last $p \times p$ diagonal block matrix being C_{uu}. Liang and Li assume C_{uu} is known but discuss that it can be estimated with partially replicated observations in practice.

Let $\rho_\tau(v) = v(\tau - 1(v < 0))$. Note that the solution to minimizing $\rho_\tau(\varepsilon_i - v)$ over $v \in \mathbb{R}$ is the τth quantile of ε_i. Liang and Li define the PQR objective function to be of the form

$$L_\tau(\beta) = \sum_{i=1}^{n} \rho_\tau\left(\frac{\hat{Y}_i - \hat{W}_i'\beta}{\sqrt{1+\beta' C_{uu}\beta}}\right) + n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \tag{8.2}$$

where $\hat{Y}_i = Y_i - \hat{m}_y(Z_i)$ and $\hat{W}_i = W_i - \hat{m}_w(Z_i)$ using the notation defined in Section 4.5, and $p_{\lambda_j}(\cdot)$ is a penalty function with tuning parameter λ_j. He and Liang (2000) propose the QR estimate of β by minimizing the first term in (8.2) and also provide insights for this. Compared with the PLS in Section 4.5, the PQR uses the factor $\sqrt{1+\beta' C_{uu}\beta}$ to correct the bias in the loss function due to the presence of measurement error in X_i. Liang and Li establish the oracle property for the PQR estimator by assuming p is fixed.

8.2 Koenker's (2011) additive models for quantile regression

Koenker (2011) considers models for conditional quantiles indexed by $\tau \in (0,1)$ of the general form

$$Q_{Y_i|X_i,Z_i}(\tau|X_i,Z_i) = X_i'\theta_0 + \sum_{j=1}^{q} g_j(Z_{ij}), \tag{8.3}$$

where X_i is a $p \times 1$ vector of regressors that enter the conditional quantile function linearly, and the nonparametric component g_i's are continuous functions, either univariate or bivariate. Let $g = (g_1, ..., g_q)'$ be a vector of functions. Koenker proposes to estimate these unknown functions and θ_0 by solving

$$\min_{(\theta,g)} \sum_{i=1}^{n} \rho_\tau \left(Y_i - X_i'\theta - \sum_{j=1}^{q} g_j(Z_{ij}) \right) + \lambda_0 \|\theta\|_1 + \sum_{j=1}^{q} \lambda_j V(\nabla g_j) \tag{8.4}$$

where $\rho_\tau(u)$ is defined as above, $\|\theta\|_1 = \sum_{k=1}^{p} |\theta_k|$ and $V(\nabla g_j)$ denotes the total variation of the derivative or gradient of the function g_j. For g with absolutely continuous derivative g', the total variation of $g' : \mathbb{R} \to \mathbb{R}$ is given by $V(g'(z)) = \int |g''(z)| dz$, while for $g : \mathbb{R}^2 \to \mathbb{R}$, $V(\nabla g) = \int \left\|\nabla^2 g(z)\right\| dz$, where $\|\cdot\|$ is the usual Hilbert-Schmidt norm for matrices and $\nabla^2 g(z)$ denotes the Hessian of $g(z)$. The Lasso penalty $\|\theta\|_1$ leads to a sparse solution for parametric component and then select the nonzero parametric components.

To select the tuning parameter λ, Koenker proposes an SIC-like criterion

$$SIC(\lambda) = n \log \hat{\sigma}(\lambda) + \frac{1}{2} p(\lambda) \log n,$$

where $\hat{\sigma}(\lambda) = n^{-1} \sum_{i=1}^{n} \rho_\tau (Y_i - \hat{g}(X_i, Z_i))$, and $p(\lambda)$ is the effective dimension of the fitted model

$$g(X_i, Z_i) = X_i'\hat{\theta} + \sum_{j=1}^{q} \hat{g}_j(Z_{ij}),$$

where Z_i is a collection of Z_{ij}'s. For linear estimator $p(\lambda)$ is defined as the trace of a pseudo projection matrix, which maps observed response into fitted values. In general form,

$$p(\lambda) = \operatorname{div}(\hat{g}) = \sum_{i=1}^{n} \frac{\partial \hat{g}(X_i, Z_i)}{\partial Y_i}.$$

He proposes some methods to obtain the pointwise and uniform confidence bands for the estimate of nonparametric components but does not study the theoretical properties of the above variable selection procedure.

8.3 Kai, Li, and Zou's (2011) composite quantile regression

Kai, Li, and Zou (2011) consider the following varying coefficient partial linear models

$$Y_i = \alpha_0(U_i) + X_i'\boldsymbol{\alpha}(U_i) + Z_i'\boldsymbol{\beta} + \varepsilon_i, \tag{8.5}$$

where (Y_i, U_i, X_i, Z_i), $i = 1, ..., n$, are i.i.d., $\alpha_0(\cdot)$ is a baseline function of scalar random variable U_i, $\boldsymbol{\alpha}(\cdot) = \{\alpha_1(\cdot), ..., \alpha_{d_1}(\cdot)\}'$ consists of d_1 unknown varying coefficient functions, $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_{d_2}\right)'$ is a

d_2-dimensional coefficient vector and ε_i is random error with zero mean and CDF $F(\cdot)$. They assume that ε_i is independent of U_i, X_i, and Z_i.

Note that the τth conditional quantile function of Y_i given $(U_i, X_i, Z_i) = (u, x, z)$ is

$$Q_\tau(u, x, z) = \alpha_0(u) + x'\boldsymbol{\alpha}(u) + z'\boldsymbol{\beta} + c_\tau$$

where $c_\tau = F^{-1}(\tau)$. All quantile regression estimates ($\hat{\boldsymbol{\alpha}}_\tau(u)$ and $\hat{\boldsymbol{\beta}}_\tau$) estimate the same target quantities ($\boldsymbol{\alpha}(u)$ and $\boldsymbol{\beta}$) with the optimal rate of convergence. Therefore, they consider combining the information across multiple quantile estimates to obtain improved estimates of $\boldsymbol{\alpha}(u)$ and $\boldsymbol{\beta}$, which leads to the *composite quantile regression* (CQR) proposed by Zou and Yuan (2008). Let $\tau_k = k/(q+1)$, $k = 1, ..., q$ for a given q. The CQR estimates of $\alpha_0(\cdot)$, $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}$ are obtained by minimizing the following CQR loss function

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \alpha_0(U_i) - X_i'\boldsymbol{\alpha}(U_i) - Z_i'\boldsymbol{\beta}\right).$$

Noting that $\alpha_j(\cdot)$ are unknown for $j = 0, 1, ..., d_1$, but they can be approximated locally by linear functions: $\alpha_j(U) \approx \alpha_j(u) + \alpha_j'(u)(U-u) = a_j + b_j(U-u)$ when U lies in the neighborhood of u. Then let $\{\tilde{\mathbf{a}}_0, \tilde{b}_0, \tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\beta}}\}$ be the minimizer of the local CQR function defined by

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\left\{Y_i - a_{0k} - b_0(U_i - u) - X_i'\left[\mathbf{a} + \mathbf{b}(U_i - u)\right] - Z_i'\beta\right\}K_h(U_i - u).$$

where $K_h(u) = K(u/h)/h$ with K and h being the kernel and bandwidth, respectively, $\mathbf{a}_0 = (a_{01}, ..., a_{0q})'$, $\mathbf{a} = (a_1, ..., a_{d_1})'$, $\mathbf{b} = (b_1, ..., b_{d_1})'$, and $\tilde{\mathbf{a}}_0 = (\tilde{a}_{01}, ..., \tilde{a}_{0q})'$, and we have suppressed the dependence of these estimates on u. Initial estimates of $\alpha_0(u)$ and $\boldsymbol{\alpha}(u)$ are then given by

$$\tilde{\alpha}_0(u) = \frac{1}{q}\sum_{k=1}^{q}\tilde{a}_{0k} \text{ and } \tilde{\boldsymbol{\alpha}}(u) = \tilde{\mathbf{a}}.$$

Given these initial estimates, the estimate of $\boldsymbol{\beta}$ can be refined by

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}}\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\left[Y_i - \tilde{a}_{0k}(U_i) - X_i'\tilde{\mathbf{a}}(U_i) - Z_i'\boldsymbol{\beta}\right]$$

which is called the semi-CQR estimator of $\boldsymbol{\beta}$. Given $\hat{\boldsymbol{\beta}}$, the estimates of the nonparametric parts can be improved by the following minimization problem

$$\min_{\mathbf{a}_0, b_0, \mathbf{a}, \mathbf{b}}\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\left[Y_i - Z_i'\hat{\boldsymbol{\beta}} - a_{0k} - b_0(U_i - u) - X_i'\left[\mathbf{a} + \mathbf{b}(U_i - u)\right]\right]K_h(U_i - u).$$

In view of the fact that variable selection is a crucial step in high-dimensional modeling, Kai, Li, and Zou focus on the selection of nonzero components in the vector $\boldsymbol{\beta}$ of parametric coefficients. Let $p_{\lambda_n}(\cdot)$ be a penalty function with tuning parameter λ_n. The penalized loss function is

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \tilde{a}_{0k}(U_i) - X_i'\tilde{\boldsymbol{\alpha}}(U_i) - Z_i'\boldsymbol{\beta}\right) + nq\sum_{j=1}^{d_2}p_{\lambda_n}\left(\left|\beta_j\right|\right).$$

Note that the objective function is nonconvex and both loss function and penalty parts are nondifferentiable. They propose to follow the one-step sparse estimate scheme in Zou and Li (2008) to derive a one-step sparse semi-CQR estimator. First, they obtain the unpenalized semi-CQR estimator $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_1^{(0)}, ..., \hat{\beta}_{d_2}^{(0)})'$. Then they define

$$G_{n,\lambda_n}(\boldsymbol{\beta}) = \sum_{k=1}^{q}\sum_{i=1}^{n} \rho_\tau \left(Y_i - \tilde{\alpha}_{0k}(U_i) - X_i'\tilde{\boldsymbol{\alpha}}(U_i) - Z_i'\boldsymbol{\beta}\right) + nq \sum_{j=1}^{d_2} p'_{\lambda_n}\left(\left|\hat{\beta}_j^{(0)}\right|\right)|\beta_j|.$$

They call $\hat{\boldsymbol{\beta}}^{OSE}(\lambda_n) = \text{argmin}_{\boldsymbol{\beta}} G_{n,\lambda_n}(\boldsymbol{\beta})$ as *one-step sparse semi-CQR estimator*.

Under some conditions, they show that $\hat{\boldsymbol{\beta}}^{OSE}$ enjoys the oracle property and the property holds for a class of concave penalties. To choose the tuning parameter λ, a BIC-like criterion is proposed as follows

$$BIC(\lambda) = \log\left[\sum_{k=1}^{q}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \hat{\alpha}_{0k}(U_i) - X_i'\hat{\boldsymbol{\alpha}}(U_i) - Z_i'\hat{\boldsymbol{\beta}}^{OSE}(\lambda)\right)\right] + \frac{\log n}{n} df_\lambda,$$

where df_λ is the number of nonzero coefficients in the parametric part of the fitted models. They propose to use $\hat{\lambda}_{BIC} = \text{argmin}_\lambda BIC(\lambda)$ as the tuning parameter.

8.4 Lin, Zhang, Bondell and Zou's (2012) sparse nonparametric quantile regression

Lin, Zhang, Bondell and Zou (2012) adopt the COSSO-type penalty to develop a new penalized framework for joint quantile estimation and variable selection. In the framework of SS-ANOVA, a function $f(x) = f\left(x^{(1)}, ..., x^{(p)}\right)$ has the ANOVA decomposition in (7.2). The entire tensor-product space for estimating $f(x)$ is given in (7.3). But in practice the higher-order interactions are usually truncated for convenience to avoid the curse of dimensionality. (7.4) gives a general expression for truncated space. Using the notation defined in Section 7.1, the regularization problem of joint variable selection and estimation is defined by

$$\min_f \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - f(X_i)\right) + \lambda \sum_{j=1}^{p} w_j \left\|P^j f\right\|_{\mathcal{F}}$$

where $P^j f$ is the projection of f on $\mathcal{F}^j$, the penalty function penalize the sum of component norms, and $w_j \in (0, \infty)$ is weight. In principle, smaller weights are assigned to important function components while larger weights are assigned to less important components. This is in the same spirit of the adaptive Lasso and adaptive COSSO. They also propose to construct the weight w_j's from the data adaptively:

$$w_j^{-1} = \left\|P^j \tilde{f}\right\|_{n,L_2} = \left\{n^{-1}\sum_{i=1}^{n}\left[P^j \tilde{f}(X_i)\right]^2\right\}^{1/2} \quad \text{for } j = 1, ..., p,$$

where $\tilde{f}$ is a reasonable initial estimator of f, say the kernel quantile regression (KQR) estimator of Li, Liu, and Zhu (2007) which is obtained by penalizing the roughness of the function estimator using

its squared functional norm in a RKHS. That is, the KQR solves the regularization problem

$$\min_{f\in\mathcal{H}_K}\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i-f\left(X_i\right)\right)+\frac{\lambda}{2}\left\|f\right\|_{\mathcal{H}_K}^2$$

where $\mathcal{H}_K$ is a RKHS and $\|\cdot\|_{\mathcal{H}_K}$ is the corresponding function norm.

An equivalent expression of the above optimization problem is

$$\min_{f,\theta}\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i-f\left(X_i\right)\right)+\lambda_0\sum_{j=1}^{p}w_j^2\theta_j^{-1}\left\|P^jf\right\|_{\mathcal{F}}^2\quad\text{s.t. }\sum_{j=1}^{p}\theta_j\leq M,\,\theta_j\geq 0$$

where both λ_0 and M are smoothing parameters. Lin, Zhang, Bondell and Zou show that the solution has the following structure

$$\hat{f}\left(x\right)=\hat{b}+\sum_{i=1}^{n}\hat{c}_i\sum_{j=1}^{p}\frac{\hat{\theta}_j}{w_j^2}R_j\left(X_i,x\right)$$

where $\hat{\mathbf{c}}=\left(\hat{c}_1,...,\hat{c}_n\right)'\in\mathbb{R}^n$, $\hat{b}\in\mathbb{R}$, and $R_j\left(X_i,x\right)$ is the reproducing kernel of subspace $\mathcal{F}^j$. Let $\mathbf{R}^\theta$ be the $n\times n$ matrix with (k,l)-th element $R_{kl}^\theta=\sum_{j=1}^{p}w_j^2\theta_j^{-1}R_j\left(X_k,X_l\right)$. Let $\mathbf{c}=\left(c_1,...,c_n\right)'$ and $\boldsymbol{\theta}=\left(\theta_1,...,\theta_p\right)'$. The objective function becomes

$$\min_{b,\mathbf{c},\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i-b-\sum_{k=1}^{n}c_kR_{kl}^\theta\right)+\lambda_0\mathbf{c}'\mathbf{R}^\theta\mathbf{c}\text{ s.t. }\sum_{j=1}^{p}\theta_j\leq M,\ \theta_j\geq 0.$$

An iterative optimization algorithm is proposed to solve the above problem.

1. Fix θ, solve $(b,\mathbf{c})$ by

$$\min_{b,\mathbf{c}}\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i-b-\sum_{k=1}^{n}c_kR_{kl}^\theta\right)+\lambda_0\mathbf{c}'\mathbf{R}^\theta\mathbf{c};$$

2. Fix $(b,\mathbf{c})$, solve θ by

$$\min_{\theta}\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i^*-\sum_{j=1}^{p}\theta_jG_{ij}\right)+\lambda_0\mathbf{c}'\mathbf{G}\boldsymbol{\theta}\text{ s.t. }\sum_{j=1}^{p}\theta_j\leq M,\ \theta_j\geq 0,$$

where $Y_i^*=Y_i-b$, and G_{ij} is the (i,j)-th element of $n\times p$ matrix $\mathbf{G}=\left(w_1^{-2}R_1\mathbf{c},...,w_p^{-2}R_p\mathbf{c}\right)$.

The optimization problems in Steps 1-2 can be cast into quadratic programming and linear programming problems, respectively. So both can be solved using standard optimization softwares. A SIC-like criterion is proposed to select the tuning parameter. However the theoretical properties of the new variable selection procedure are not discussed.

9 Concluding Remarks

In this chapter we survey some of the recent developments on variable selections in nonparametric and semiparametric regression models. We focus on the use of Lasso, SCAD, or COSSO-type penalty

for variable or component selections because of the oracle property of the SCAD and the adaptive versions of Lasso and COSSO. The oracle property has been demonstrated for some of the variable selection procedures but not for others (e.g., variable selection in nonparametric/semiparametric QR). It is interesting to develop variable selection procedures with the oracle property for some of the models reviewed in this chapter. In addition, the i.i.d. assumption has been imposed in almost all papers in the literature. It is important to relax this assumption to allow for either heterogeneity or serial/spatial dependence in the data. More generally, one can study variable selection for more complicated semiparametric/nonparametric models via shrinkage.

Despite the huge literature on Lasso or SCAD type techniques in statistics, we have seen very few developments of them in econometrics until 2009. Almost all of the works on variable selection in statistics are based on the assumption that the regressors are uncorrelated with or independent of the error terms, namely, they are exogenous. However, in economic applications there are many examples in which some covariates are endogenous due to measurement error, omitted variables, sample selection, or simultaneity. The endogeneity causes inconsistent estimate by the PLS method and misleading statistical inference and one has to resort to instrumental variables (IVs) to handle this problem. Caner (2009) seems to be the first published paper to address this issue through shrinkage GMM estimation. Since then we have observed a large literature on the use of Lasso or SCAD type techniques in econometrics to cope with endogeneity in parametric models. They fall into three categories. The first category focuses on selection of covariates or parameters in the structural equation, see Caner (2009), Caner and Zhang (2009), and Fan and Liao (2011, 2012). Caner (2009) considers covariate selection in GMM with Bridge penalty when the number of parameters is fixed; Caner and Zhang (2009) study covariate selection in GMM via adaptive elastic-net estimation by allowing the number of parameters to diverge to infinity; Fan and Liao (2011) consider variable selection with endogenous covariates in ultra high-dimensional regressions via penalized GMM and penalized empirical likelihood (EL); Fan and Liao (2012) propose a penalized focused GMM (FGMM) criterion function to select covariates. The second category focuses on the selection relevant IVs (or deletion of irrelevant/weak IVs); see Belloni, Chernozhukov, and Hansen (2010), Caner and Fan (2011), and García (2011). Belloni, Chernozhukov, and Hansen (2010) introduce a heteroskedasticity-consistent Lasso-type estimator to pick optimal instruments among many of them. Caner and Fan (2011) use the adaptive Lasso to distinguish relevant and irrelevant/weak instruments in heteroskedastic linear regression models with fixed numbers of covariates and IVs. García (2011) proposes a two stage least squares (2SLS) estimator in the presence of many weak and irrelevant instruments and heteroskedasticity. The third category focuses on the selection both covariates and valid IVs; see Liao (2011) and Gautier and Tsybakov (2011). Liao (2011) considers the selection of valid moment restrictions via adaptive Lasso, Bridge and SCAD, and the selection of group variables and group valid moment restrictions via adaptive group Lasso when the number of parameters is fixed. Gautier and Tsybakov (2011) extend the Dantzig selector of Candès and Tao (2007) to the linear GMM framework and propose a new procedure called *self tuning instrumental variable* (STIV) estimator for the selection of covariates and valid IVs when the number of covariates/parameters can be larger than the sample size. Nevertheless, none of these works address the issue of flexible functional form. It is interesting to consider variable selection for

semiparametric or nonparametric models with endogeneity. Sieve estimation or local GMM estimation via shrinkage seems to be a very promising field to delve into.

References

Avalos, M., Grandvalet, Y., and Ambroise, C. 2007. "Parsimonious Additive Models." *Computational Statistics & Data Analysis* 51, pp.2851-2870.

Bach, F.R. 2008. "Consistency of the Group Lasso and Multiple Kernel Learning." *Journal of Machine Learning Research* 9, pp.1179-1225.

Belloni, A., Chernozhukov, V., and Hansen, C. 2010. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Working Paper*, MIT.

Belloni, A., and Chernozhukov, V. 2011a. "l_1-penalized Quantile Regression in High-dimensional Sparse Models." *Annals of Statistics* 39, pp.82-130.

Belloni, A., and Chernozhukov, V. 2011b. "High-Dimensional Sparse Econometric Models: An Introduction." In: *Inverse Problems and High Dimensional Estimation*, Alquier, P., Gautier, E., and Stoltz, G., eds. *Lectures in Statistics* 203, pp.127-162. Springer, Berlin.

Bernanke, B. S., Bovin, J., and Eliasz, P. 2005. "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach." *Quarterly Journal of Economics* 120, pp.387-422.

Bertin, K., and Lecué, G. 2008. "Selection of Variable and Dimension Reduction in High-dimensional Non-parametric Regression." *Electronic Journal of Statistics* 2, pp.1223-1241.

Bunea, F. 2004. "Consistent Covariate Selection and Post Model Selection Inference in Semiparametric Regression." *Annals of Statistics* 32, pp.898-927.

Bunea, F. 2008. "Consistent Selection via the Lasso for High Dimensional Approximating Regression Models." *Institute of Mathematical Statistics Collection* 3, pp.122-137.

Comminges, L., and Dalayan, A.S. 2011. "Tight Condition for Consistent Variable Selection in High Dimensional Nonparametric Regression." *JMLR: Workshop and Conference Proceedings* 19, pp.187-205.

Caner, M. 2009. "Lasso-type GMM Estimator." *Econometric Theory* 25, pp.270-290.

Caner, M., and Fan, M. 2011. "A Near Minimax Risk Bound: Adaptive Lasso with Heteroskedastic Data in Instrumental Variable Selection." *Working Paper*, North Carolina State University.

Caner, M., and Zhang, H. H. 2009. "General Estimating Equations: Model Selection and Estimation with Diverging Number of Parameters." *Working Paper*, North Carolina State University.

Candès, E.J., and Tao, T. 2007. "The Dantzig Selector: Statistical Estimation When p is Much Larger than n." *Annals of Statistics* 35, pp.2313-2351.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P., 1997. "Generalized Partially Linear Single-index Models." *Journal of American Statistical Association* 92, pp.477-489.

Chen, B., Yu, Y., Zou, H. and Liang, H. 2012. "Profiled Adaptive Elastic-Net Procedure for Partially Linear Models." *Journal of Statistical Planning and Inference* 142, pp.1773-1745.

Donoho, D. L., and Johnstone, I. M. 1994. "Ideal Spatial Adaptation via Wavelet Shrinkages." *Biometrika* 81, pp.425-455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. 1995. "Wavelet Shrinkage: Asymptopia (with discussion)?" *Journal of the Royal Statistical Society, Series B* 57, pp.301-369.

Efron B., Hastie, T., Johnstone, I., and Tibshirani, R. 2004. "Least Angle Regression." *Annals of Statistics* 32, pp.407-499.

Fan, J., Feng, Y. and Song, R. 2011. "Nonparametric Independence Screening in Sparse Ultra-high Dimensional Additive Models." *Journal of American Statistical Association* 116, pp.544-557.

Fan, J., Huang, T. and Peng, H. 2005. "Semilinear High-dimensional Model for Normalization of Microarray Data: a Theoretical Analysis and Partial Consistency (with discussion)." *Journal of American Statistical Association* 100, pp.781-813.

Fan, J., and Li., R. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96, pp.1348-1360.

Fan, J., and Li., R. 2004. "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis." *Journal of the American Statistical Association* 99, pp.710-723.

Fan, J., and Li., R. 2006. "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Recovery." *Proceedings of the International Congress of Mathematicians, Madrid, Spain,* pp.595-622.

Fan, J., and Liao, Y. 2011. "Ultra High Dimensional Variable Selection with Endogenous Covariates." *Working Paper*, Princeton University.

Fan, J., and Liao, Y. 2012. "Endogeneity in Ultrahigh Dimension." *Working Paper*, Princeton University.

Fan, J., and Lv, J. 2010. "A Selection Overview of Variable Selection in High Dimensional Feature Space." *Statistica Sinica* 20, pp.101-148.

Fan, J., Lv, J., and Qi, L. 2011. "Sparse High-dimensional Models in Economics." *Annual Review of Economics* 3, pp.291-317.

Fan, J., and Peng, H. 2004. "On Non-Concave Penalized Likelihood With Diverging Number of Parameters." *Annals of Statistics* 32, pp.928-961.

Fan, J., Yao, Q., and Cai, Z. 2003. "Adaptive Varying-coefficient Linear Models." *Journal of the Royal Statistical Society, Series B* 65, pp.57-80.

Fan, J., and Zhang, J.T. 1998. "Functional Linear Models for Longitudinal Data." *Journal of the Royal Statistical Society, Series B* 39, pp.254-261.

Fan J., Zhang, J., and Yu, K. 2011. "Asset Allocation and Risk assessment with Gross Exposure Constraints for vast Portfolios." *Working Paper*, Princeton University.

Frank, I.E., and Friedman, J.H. 1993. "A Statistical View of Some Chemometrics Regression Tools (with discussion)." *Technometrics* 35, pp.109-148.

Friedman, J., Hastie, T., and Tibshirani, R. 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics* 9, pp.432-441.

Fu, W. 1998. "Penalized Regressions: The Bridge versus the Lasso." *Journal of Computational and Graphical Statistics* 7, pp.397-416.

García, P. E. 2011. "Instrumental Variable Estimation and Selection with Many Weak and Irrelevant Instruments." *Working Paper*, University of Wisconsin, Madison.

Green, P. J., and Silverman, B. W. 1994. *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Härdle, W., and Stoker, T.M. 1989. "Investigating Smooth Multiple Regression by the Method of Average Derivatives." *Journal of American Statistical Association* 84, pp.986-995.

He, X., and Liang H. 2000. "Quantile Regression Estimates for a Class of Linear and Partially Linear Errors-in-Variables Models." *Statistica Sinica* 10, pp.129-140.

Horel, A.E., and Kennard, R.W. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12, pp.55-67.

Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. 2001. "Structure Adaptive Approach for Dimension Reduction." *Annals of Statistics* 29, pp.1537-1566.

Huang, J. Breheny P., and Ma, S. 2012. "A Selective Review of Group Selection in High Dimensional Models". Forthcoming in *Statistical Science.*

Huang, J., Horowitz, J.L., and Wei, F. 2010. "Variable Selection in Nonparametric Additive Models." *Annals of Statistics* 38, pp.2282-2313.

Huang, J.Z., Wu, C.O., and Zhou, L. 2002. "Varying-coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements." *Biometrika* 89, pp.111-128.

Hunter, D.R., and Li. R. 2004. "Variable Selection Using MM Algorithms." *Annals of Statistics* 33, pp.1617-1642.

Ichimura, H. 1993. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models." *Journal of Econometrics* 58, pp.71-120.

Jagannathan, R., and Ma, T. 2003. "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *Journal of Finance* 58, pp.1651-1683.

Kai, B., Li, R., and Zou, H. 2011. "New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models." *Annals of Statistics* 39, pp.305-332.

Kato R., and Shiohama, T. 2009. "Model and Variable Selection Procedures for Semiparametric Time Series Regression." *Journal of Probability and Statistics,* Article ID 487194, 37 pages..

Knight, K. and Fu, W. 2000. "Asymptotics for Lasso-type Estimators." *Annals of Statistics* 28, pp.1356-1378.

Koenker, R. 2004. "Quantile Regression for Longitudinal Data." *Journal of Multivariate Analysis* 91, pp.74-89.

Koenker, R. 2005. *Quantile Regression.* Cambridge University Press, Cambridge.

Koenker, R. 2011. "Additive Models for Quantile Regression: Model Selection and Confidence Bands." *Brazilian Journal of Probability and Statistics* 25, pp.239-262.

Kong, E., and Xia, Y. 2007. "Variable Selection for the Single-index Model." *Biometrika* 94, pp.217-229.

Lafferty, J., and Wasserman, L. 2008. "Redeo: Sparse, Greedy Nonparametric Regression." *Annals of Statistics* 36, pp.18-63.

Leng, C., and Zhang, H. H. 2006. "Model Selection in Nonparametric Hazard Regression." *Nonparametric Statistics* 18, pp.316-342.

Li, K. C. 1991. "Sliced Inverse Regression for Dimensional Reduction (with discussion)." *Journal of the American Statistical Association* 86, pp.417-429.

Li, K. C., Duan, N. H. 1989. "Regression Analysis under Link Violation." *Annals of Statistics* 17, pp.1009-1052.

Li, R., and Liang, H. 2008. "Variable Selection in Semiparametric Regression Modeling." *Annals of Statistics* 36, pp.261-286.

Li, Y., Liu, Y., and Zhu, J. 2007. "Quantile Regression in Reproducing Kernel Hilbert Spaces." *Journal of the American Statistical Association* 102, pp.255-267.

Lian, H. 2010. "Flexible Shrinkage Estimation in High-Dimensional Varying Coefficient Models." *Working Paper*, NTU.

Liang, H., and Li, R. 2009. "Variable Selection for Partially Linear Models with Measurement Errors." *Journal of the American Statistical Association* 104, pp.234-248.

Liang, H., Liu, X., Li, R., and Tsai, C.-L. 2010. "Estimation and Testing for Partially Linear Single-index Models." *Annals of Statistics* 38, pp.3811-3836.

Liang, H., and Wang, N. 2005. "Partially Linear Single-index Measurement Error Models." *Statistica Sinica* 15, pp.99-116.

Liao, Z. 2011. "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection." *Working Paper*, UCLA.

Lin C., Zhang, H.H., Bondell, H.D., and Zou, H. 2012. "Variable Selection for Nonparametric Quantile Regression via Smoothing Spline ANOVA". *Working Paper*, North Carolina State University.

Lin, Y., and Zhang, H.H. 2006. "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models-COSSO." *Annals of Statistics* 34, pp.2272-2297.

Liu, X., Wang, L., and Liang, H. 2011. "Estimation and Variable Selection for Semiparametric Additive Partially Linear Models." *Statistica Sinica* 21, pp.1225-1248.

Meier, L., Van De Geer, S.A., and Bühlmann, P. 2009. "High-dimensional Additive Modeling." *Annals of Statistics* 37, pp.3379-3821.

Miller, H., and Hall, P. 2010. "Local Polynomial Regression and Variable Selection." Borrowing Strength: Theory Powering Applications-A Festschrift for Lawrence D. Brown. *IMS Collections* Vol. 6, pp.216-233.

Naik, P. A., and Tsai, C.-L. 2001. "Single-index Model Selections." *Biometrika* 88, pp.821-832.

Ni, X., Zhang, H. H., and Zhang, D., 2009. "Automatic Model Selection for Partially Linear Models." *Journal of Multivariate Analysis* 100, pp.2100-2111.

Peng, H., and Huang, T. 2011. "Penalized Least Squares for Single Index Models." *Journal of Statistical Planning and Inference* 141, pp.1362-1379.

Qu, A., and Li, R. 2006. "Quadratic Inference Functions for Varying Coefficient Models with Longitudinal Data." *Biometrics* 62, pp.379-391.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. 2008. "Spam: Sparse Additive Models." *Advances in Neural Information Processing Systems* 20, pp.1202-1208.

Rinaldo, A., 2009. "Properties and Refinement of the Fused Lasso." *Annals of Statistics* 37, pp.2922-2952.

Storlie, C.B., Bondell, H.D., Reich, B.J., and Zhang, H.H. 2011. "Surface Estimation, Variable Selection, and the Nonparametric Oracle Property." *Statistica Sinica* 21, pp.679-705.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B* 58, pp.267-288.

Tibshirani, R. 2011. "Regression Shrinkage and Selection via the Lasso: a Retrospective." *Journal of the Royal Statistical Society, Series B* 73, pp.273-282.

Tibshirani, R.J., Hoefling, H., and Tibshirani, R. 2010. "Nearly-Isotonic Regression." *Working paper*, Stanford University.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. 2005. "Sparsity and Smoothness via the Fused Lasso." *Journal of the Royal Statistical Society, Series B* 67, pp.91-108.

Wahba, G., 1990. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia, PA.

Wang, H., and Leng, C. 2008 "A Note of Adaptive Group Lasso." *Computational Statistics and Data Analysis* 52, pp.5277-5286.

Wang, H., Li, G., and Jiang, G. 2007. "Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso." *Journal of Business & Economic Statistics* 25, pp.347-355.

Wang, H., and Xia, Y. 2009 "Shrinkage Estimation of the Varying Coefficient Model." *Journal of the American Statistical Association* 104, pp.747-757.

Wang, L., Chen, G., and Li, H. 2007. "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data." *Bioinformatics* 23, pp.1486-1494.

Wang, L., Li, H., and Huang, J.H. 2008. "Variable Selection in Nonparametric Varying-coefficient Models for Analysis of Repeated Measurements." *Journal of the American Statistical Association* 103, pp.1556-1569.

Wang, T., Xu P.-R., and Zhu L.-X. 2012. "Non-convex Penalized Estimation in High-dimensional Models with Single-index Structure." *Journal of Multivariate Analysis* 109, pp.221-235.

Wei, F., and Huang, J. 2010. "Consistent Group Selection in High-dimensional Linear Regression." *Bernoulli* 16, pp.1369-1384.

Wei, X., Huang, J., and Li, H. 2011. "Variable Selection and Estimation in High-dimensional Varying-coefficient Models." *Statistica Sinica* 21, pp.1515-1540.

Wu, Y., and Liu, Y. 2009. "Variable Selection in Quantile Regression." *Statistic Sinica* 19, pp.801-817.

Xia, Y., Tong, H., Li, W.K., and Zhu, L. 2002. "An Adaptive Estimation of Dimension Reduction Space." *Journal of the Royal Statistical Society, Series B* 64, pp.363-410.

Xie, H., and Huang, J. 2009. "SCAD-penalized Regression in High-dimensional Partially Linear Models." *Annals of Statistics* 37, pp.673-696.

Xue, L. 2009. "Consistent Variable Selection in Additive Models." *Statistica Sinica* 19, pp.1281-1296.

Yang, B. 2012. *Variable Selection for Functional Index Coefficient Models and Its Application in Finance and Engineering*. Ph.D. thesis, University of North Carolina at Charlotte.

Yuan, M., and Lin, Y. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society, Series B* 68, pp.49-67.

Yuan, M., and Lin, Y. 2007. "Model Selection and Estimation in the Gaussian Graphical Model." *Biometrika* 94, pp.19-35.

Zeng, P., He, T., and Zhu, Y. 2011. "A Lasso-type Approach for Estimation and Variable Selection in Single Index Models." *Journal of Computational and Graphical Statistical* 21, pp.92-109.

Zhang, C.-H. 2010. "Nearly Unbiased Variable Selection under Minimax Concave Penalty." *Annals of Statistics* 38, pp.894-932.

Zhang, H., and Lin, Y. 2006. "Component Selection and Smoothing for Nonparametric Regression in Exponential Families." *Statistica Sinica* 16, pp.1021-1042.

Zhao, P. and Xue, L. 2011. "Variable Selection for Varying Coefficient Models with Measurement Errors." *Metrika* 74, pp.231-245.

Zhao, P. and Yu, B. 2006. "On Model Selection Consistency of Lasso." *Journal of Machine Learning Research* 7, pp.2541-2563.

Zhu, L.-P., Qian, L., and Lin, J. 2011. "Variable Selection in a Class of Single-index Models." *Annals of the Institute of Statistical Mathematics* 63, pp.1277-1293.

Zhu, L.-P., and Zhu, L.-X. 2009. "Nonconcave Penalized Inverse Regression in Single-index Models with High Dimensional Predictors." *Journal of Multivariate Analysis* 100, pp.862-875.

Zou, H. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101, pp.1418-1429.

Zou, H., and Hastie, T. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B* 67, pp.301-320.

Zou, H., and Li, R. 2008. "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models (with discussion)." *Annals of Statistics* 36, pp.1509-1533.

Zou, H., and Yuan, M. 2008. "Composite Quantile Regression and the Oracle Model Selection Theory." *Annals of Statistics* 36, pp.1509-1533.

Zou, H., and Zhang, H.H. 2009. "On the Adaptive Elastic-net with a Diverging Number of Parameters." *Annals of Statistics* 37, pp.1733-1751.