ON SEMI-PARAMETRIC MODEL AND SUBSET SELECTION

KONG EFANG

(M.Sc., Beijing Normal University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2006

Acknowledgements

I want to express my gratitude to my supervisor Assoc Professor Xia Yingcun and Chen Songxi. Being great researchers themselves, they inspired my interest in nonparametric statistics and their many valuable suggestions have greatly influenced the direction of my research. Prof Xia, in particular has provided an encouraging yet critical atmosphere for my research and has guided me throughout my study in NUS. This thesis would have been impossible without discussions with and insight from them.

At different stages of my stay at NUS, I received help from the academic or the secretarial staff in the Department of Statistics and Applied Probability. I'm really grateful to all of them.

I would like to contribute the completion of this thesis to my dearest family, who have always been supporting me with incredible encouragement and understanding all these years.

Contents

Summary

As a gray area between parametric and nonparametric methods, semi-parametric model has generated a large literature in econometrics and statistics. By semi-parametric approach, we mean models and estimation problems that involve an unknown smooth function and a finite number of unknown parameters. By relaxing the rigid assumption imposed on the form of the functional by parametric methods, such as linear or polynomial, semi-parametric approach allows for more flexible modeling, while avoiding the 'curse of dimensionality' suffered by nonparametric models since the unknown function is defined in one dimension space. There have been quite a few recent monographs on this topic ([6, 7, 64]) and it is shown that semi-parametric techniques have indeed much to offer in practice. In this thesis, two aspects of application of semi-parametric models are discussed: subset selection and financial time series modeling.

Subset selection has always been a critical and challenging issue in regression analysis. Exclusion of irrelevant variables not only delivers parsimonious models which facilitate explanation, but also improves estimation precision and forecasting accuracy. In linear regression models, it is well-known that the leave-one-out cross-validation is inconsistent, while the leave-m-out cross-validation(CV(m)) is ([73]). But the Balanced Incomplete Block Design assumption necessitated by CV(m) is not easy to verify in practice. Motivated by the properties of cross-validation methods under nonparametric settings, a new consistent method based on semi-parameterization is proposed in Chapter 1. Simulations show that this approach has very good finite sample performance, which is further backed up by applications to a pollution data set.

In the second chapter, subset selection issue in the single-index model, which is a type of semi-parametric model, is discussed. I prove that CV(m) behaves differently in the single-index model from in linear regression models or in nonparametric regression models. A new consistent selection algorithm, called the separated cross-validation (SCV), is proposed. Further analysis suggests that this method has robust finite sample performance and is computationally easier than CV(m). SCV applied to the Swiss banknotes data and the ozone concentration data, leads to single-index models with selected variables that have better prediction capability than models based on all the covariates.

The last chapter focuses on financial time series modeling in which respect, the ARCH and GARCH models are among the most powerful tools in depicting the volatility clustering phenomena. However, due to the time homogeneous structure, neither ARCH nor GARCH is capable of grasping the time varying characteristics exhibited by most financial data over long time spans. As an integration of the ARCH model and the monotone varying coefficient model, the newly introduced model inherits the flexibility of varying coefficient models, while preserving the additive structure of the ARCH model. Its estimation and theoretical property are discussed. Simulation results and real data analysis are also available.

List of Tables

List of Figures

Chapter 1

Subset Selection for Linear Regression Models

1.1 Introduction

Due to its simplicity, linear regression model has been one of the most widely used and fully investigated models. Although as many covariates as possible can be taken into modeling, unnecessarily large models not only lead to estimation insufficiency but also result in difficulty for model explanation. A lot of work has been done in the literature. Examples are AIC ([1, 75]); the C_p method ([50]); BIC ([33]); the final prediction error(FPE) method ([76]); the generalized information criterion ([68]); the leave-one-out cross-validation(CV) method ([80]); the generalized cross-validation(GCV) method ([16]); the v-fold cross-validation method ([10]) and the bootstrap model selection method ([19, 20, 74]). More recent work includes [28, 81].

So far, the classical CV method and its variations (e.g. GCV) or equivalents (e.g. AIC) have been the main focus of researchers' attention in model identification and variable selection. A good survey can be found in [57, 58]. However, it is proved that for linear regression models, both CV and AIC are conservative, as they have an inclination for unnecessarily large models. Several modifications have been made on the AIC method, which is defined as $n \log(\hat{\sigma}) + c_n p$, where $\hat{\sigma}$ is the mean of the residual squares of the working model, p is the number of covariates and $c_n = 2$. In AIC, $c_n p$ can be regarded as a penalty against choosing too large a p. The basic idea of modification is to increase the penalty against including too many covariates. Well-known modifications are $c_n = \log(n)$ ([71]) and $c_n = c \log\log(n)$ ([33]). Another modification called 'leave-m-out' CV, denoted by CV(m) herein, increases the penalty by each time leaving m observations out as the test set. [73]proved that if $n - m \to \infty$ and $m/n \to 1$, then CV(m) is consistent. However, his findings are based on the Balanced Incomplete Block Design (BIBD) assumption that the sample covariance matrix of any test set of size m is asymptotically uniformly invariant as shown in (1.5), which is usually not easy to justify in practice.

While in nonparametric settings, the CV method is consistent due to the 'heavier penalty' mechanism resulted from kernel smoothing; see, e.g. [13, 82]. Motivated by the above facts, it is promising to address the subset selection issue in linear regression models by *semi-parameterization*, i.e. we treat linear regression models as semi-parametric models. We will show that this method is consistent. Compared with CV(m), it is easy to implement and is robust against the choice of the smoothing parameter.

1.2 Optimal Model and Review of Cross-validation

Consider the linear model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{1.1}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^\top$, $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^\top$ with $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2 I$, $\mathbf{X} = (X_1, \cdots, X_p)$ is a $n \times p$ matrix, and $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^\top$ is an unknown parameter. We adopt the notations in [73]. As some of the components of β in model (1.1) may be zero, a more compact model might be

$$y = \mathbf{x}_\alpha^\top \beta_\alpha + \epsilon, \tag{1.2}$$

where α is a subset of d_α distinct positive integers that are less or equal to p and $\mathbf{x}_\alpha$ (or β_α) is the d_α dimensional vector which consists of the components of $\mathbf{x}$ (or β) which are indexed by the integers in α. Let $\mathcal{A}$ denote all nonempty subsets of $\{1, \cdots, p\}$. Nominally, there are $2^p - 1$ possible different models of the form (1.2), each of which corresponds to a subset α and is denoted by $\mathcal{M}_\alpha$. The size of $\mathcal{M}_\alpha$ is defined to be d_α, the number of predictors in $\mathcal{M}_\alpha$. Under model $\mathcal{M}_\alpha$, the least squares estimator of β_α is

$$\hat{\beta}_\alpha = \left(\mathbf{X}'_\alpha \mathbf{X}_\alpha\right)^{\mathbf{-1}} \mathbf{X}_\alpha^\top \mathbf{y} \tag{1.3}$$

where $\mathbf{X}_\alpha = (\mathbf{x_{1\alpha}}, \cdots, \mathbf{x_{n\alpha}})^\top$ is an $n \times d_\alpha$ matrix assumed of full column rank for any $\alpha \in \mathcal{A}$, and $\mathbf{x_{i\alpha}}$ is the d_α dimensional vector containing the components of $\mathbf{x_i}$ that are indexed by the integers in α. If we know exactly which components of β are zeros, all candidate models $\{\mathcal{M}_\alpha : \alpha \subseteq \{1, \cdots, p\}\}$ can be classified into two categories:

- Category 1: at least one nonzero component of β is missing in β_α.

- Category 2: β_α contains all nonzero components of β.

Obviously, models in Category 1 are incorrect and those in Category 2 may contain redundant variables. The true model, denoted by $\mathcal{M}_{\alpha_0}$, is the one in Category 2 with the smallest size d_0 .

Leave-m-out cross-validation method CV(m) selects a model which among all $\mathcal{M}_\alpha$ minimizes the estimated squared prediction error. First, we split the data into two sets: test set $\{(y_i, \mathbf{x_i}), i \in s\}$ and learning set $\{(y_i, \mathbf{x_i}), i \in s^c\}$, where s is a subset of $\{1, \cdots, n\}$ containing m integers and s^c is its complement containing $n - m$ integers. The model $\mathcal{M}_\alpha$ is fitted from the learning set and the prediction error is assessed using the test set, treated as if they were future values. The average squared prediction error is defined as

$$m^{-1}\|\mathbf{y}_s - \mathbf{X}_{s,\alpha}\hat{\beta}_\alpha^{\setminus s}\|^2, \tag{1.4}$$

where $\|\mathbf{a}\| = (\mathbf{a}^\top \mathbf{a})^{1/2}$ for a vector $\mathbf{a}$, $\mathbf{X}_{s,\alpha}$ is the $m \times d_\alpha$ matrix containing the rows of $\mathbf{X}_\alpha$ indexed by $i \in s$, y_s is the m dimensional vector consisting of the components of $\mathbf{y}$ indexed by $i \in s$ and $\hat{\beta}_\alpha^{\setminus s}$ is the least square estimator of β_α from the learning set.

Suppose $\mathcal{B}$ is a selected collection of b size m subsets of $\{1, \cdots, n\}$, which satisfies the Balanced Incomplete Block Design([73])

$$\sup_{m\to\infty} \max_{s\in\mathcal{B}} \|\frac{1}{m}\sum_{i\in s} \mathbf{x}_i\mathbf{x}_i^\top - \frac{1}{n-m}\sum_{i\notin s} \mathbf{x}_i\mathbf{x}_i^\top\| = 0. \tag{1.5}$$

For each model $\mathcal{M}_\alpha$, the cross-validation estimate of prediction error, denoted by CV$_\alpha(m)$ is obtained by averaging (1.4) over $\mathcal{B}$, i.e.

$$CV_\alpha(m) = \frac{1}{bm}\sum_{s\in\mathcal{B}} \|\mathbf{y}_s - \mathbf{X}_{s,\alpha}\hat{\beta}_\alpha^{\setminus s}\|^2, \tag{1.6}$$

and the model with the smallest value of $\mathrm{CV}_\alpha(m)$ is the preferred model.

We here give the asymptotic expansion for $CV_\alpha(1)$ and $CV_\alpha(m)$. Let $P_\alpha = \mathbf{X}_\alpha \left(\mathbf{X}_\alpha^\top \mathbf{X}_\alpha\right)^{-1} \mathbf{X}_\alpha^\top$, and $\Delta_{\alpha,n} = n^{-1}\beta^\top \mathbf{X}^\top (I_n - P_\alpha)\mathbf{X}\beta$. Suppose (1.5) hold and

$$\mathbf{X}^\top \mathbf{X} = O(n), \quad (\mathbf{X}^\top \mathbf{X})^{-1} = O(n^{-1}), \text{ and } \lim_{n\to\infty} \max_{i\leq n} p_{i\alpha} = 0, \ \forall \alpha \in \mathcal{A}, \tag{1.7}$$

where $p_{i\alpha}$ is the ith diagonal element of the projection matrix P_α. [73] proved that

$$CV_\alpha(1) = \sigma^2 + \frac{1}{n} d_\alpha \sigma^2 + \Delta_{\alpha,n} + o_p(1), \text{ if } \mathcal{M}_\alpha \text{ is in Category 1 ,} \tag{1.8}$$

$$CV_\alpha(1) = \frac{1}{n}\epsilon^\top \epsilon + \frac{2}{n} d_\alpha \sigma^2 - \frac{1}{n}\epsilon^\top P_\alpha \epsilon + o_p(1), \text{ if } \mathcal{M}_\alpha \text{ is in Category 2.} \tag{1.9}$$

Therefore, based on (1.8) and (1.9), if

$$\liminf_{n\to\infty} \Delta_{\alpha,n} > 0, \qquad \text{for any } \mathcal{M}_\alpha \text{ in Category 1,} \tag{1.10}$$

then the chance for $CV(1)$ to eliminate useful variables tends to zero, while the probability of taking in extra variables does not. Specifically, for any $\mathcal{M}_\alpha$ with $\alpha \supset \alpha_0$,

$$P\{\mathcal{M}_\alpha \text{ is preferred to } \mathcal{M}_{\alpha_0} \text{ by } CV(1)\} = P\{2\delta_d \sigma^2 < \epsilon^\top (P_\alpha - P_{\alpha_0})\epsilon\} + o(1).$$

where $\delta_d := d_\alpha - d_0$. If ϵ is distributed as $N(0, \sigma^2 I_n)$, then as $n \to \infty$,

$$P\{\mathcal{M}_\alpha \text{ is preferred to } \mathcal{M}_{\alpha_0} \text{ by } CV(1)\} \to P\{2\delta_d < \chi^2(\delta_d)\} \neq 0, \tag{1.11}$$

where $\chi^2(\delta_d)$ is the chi-square random variable with δ_d degrees of freedom.

$CV(m)$ method rectified this inconsistency by providing more accurate assessment of the prediction error as more observations are used for validation. [73] showed that if a subset collection $\mathcal{B}$ satisfies (1.5) with $n - m \to \infty$ and $m/n \to 1$, then

$$CV_\alpha(m) = n^{-1}\epsilon^\top \epsilon + (n-m)^{-1} d_\alpha \sigma^2 + o_p((n-m)^{-1}), \text{ if } \mathcal{M}_\alpha \text{ is in Category 2.}$$

However, there are some disadvantages about the $CV(m)$ method. Firstly, it is difficult to verify whether there exists such a subset collection satisfying (1.5). Secondly, even if it does exist, the computational workload is formidable, since $(2^p - 1)$ different models need to be evaluated, if the full covariate set contains p variables.

1.3 Variable Selection by Separation

Note that if δ_d, the difference in the numbers of parameters to be estimated under model M_α and M_{α_0}, tends to infinity, the probability in (1.11) will tend to 0. This implies that if we can 'force' the unnecessary large model into one with 'infinite number of parameters', then the consistency property can be materialized. To this end, first note that any linear model M_α with $\alpha_1 = \alpha \cup k \supseteq \alpha_0$ is a special case of the partially linear model([69, 78])

$$y = \mathbf{x}_\alpha' \beta_\alpha + g(z) + \epsilon, \quad z = \mathbf{x}_k \tag{1.12}$$

with $g(z)$ set to be $\beta_k z$, where β_k is the kth component of β. The 'forced' presence of an unknown function $g(.)$ means that the number of parameters in (1.12) and thus δ_d in (1.11) is infinite.

The estimation of (1.12) were studied by[69, 78]. Note that $E(y|z) = \beta^\top E(x|z) + g(z)$, $y - E(y|z) = \beta^\top \{\mathbf{x} - E(\mathbf{x}|z)\} + \epsilon$, which suggests that estimates of the regression functions $E(y|z)$ and $E(\mathbf{x}|z)$ be inserted prior to application of the no-intercept ordinary least square method. While a variety of nonparametric estimators is available, we here consider Nadaraya-Waston estimator for $E(\mathbf{x}|z)$ and $E(y|z)$. Let

$$\tilde{\mathbf{x}}_\alpha(z) = \frac{\sum_{i=1}^n K_h(z_i - z)\mathbf{x}_{i\alpha}}{\sum_{i=1}^n K_h(z_i - z)}, \qquad \tilde{y}(z) = \frac{\sum_{i=1}^n K_h(z_i - z) y_i}{\sum_{i=1}^n K_h(z_i - z)},$$

where $K(.)$ is a symmetric univariate density function, $h > 0$ is a bandwidth and $K_h(.) = K(./h)$. Then β_α in (1.12) can be estimated by

$$\hat{\beta}_\alpha = \Big[\sum_{i=1}^{n}\{\mathbf{x}_{i\alpha} - \tilde{\mathbf{x}}_\alpha(z_i)\}\{\mathbf{x}_{i\alpha} - \tilde{\mathbf{x}}_\alpha(z_i)\}^\top\Big]^{-1}\sum_{i=1}^{n}\{\mathbf{x}_{i\alpha} - \tilde{\mathbf{x}}_\alpha(z_i)\}\{y_i - \tilde{y}(z_i)\},$$

Under some regularity conditions, [69] proved that $\hat{\beta}_\alpha$ is root-n consistent. Let

$$\tilde{\mathbf{x}}_\alpha^{\backslash i}(z) = \sum_{l \neq i} K_h(z_l - z)\mathbf{x}_l \Big/ \sum_{l \neq i} K_h(z_l - z),\ \tilde{y}^{\backslash i}(z) = \sum_{l \neq i} K_h(z_l - z)y_l \Big/ \sum_{l \neq i} K_h(z_l - z),$$
$$\hat{g}^{\backslash i}(z) = \tilde{y}^{\backslash i}(z) - \tilde{\mathbf{x}}_\alpha^{\backslash i}(z)^\top \hat{\beta}_\alpha,\ \hat{y}_{l\alpha}^{\backslash i} = \hat{g}^{\backslash i}(z_l) + \mathbf{x}_{l\alpha}^\top \hat{\beta}_\alpha.$$

Then the average prediction error is defined as

$$SCV(\alpha, z) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i\alpha}^{\backslash i})^2 = \frac{1}{n}\sum_{i=1}^{n}\Big[y_i - \tilde{y}^{\backslash i}(z_i) - \{\mathbf{x}_{i\alpha} - \tilde{\mathbf{x}}_\alpha(z_i)\}^\top \hat{\beta}_\alpha\Big]^2. \quad (1.13)$$

The selection algorithm goes as follows. Start with an initial set $\alpha = \{i_1, \cdots, i_d\} \supseteq \alpha_0$.

Step 1. Compute $CV_{\alpha_k}(1)$ for every $\alpha_k = \alpha \setminus \{i_k\}$, $k = 1, \cdots, d$ and $k := \min_{1 \le j \le d} CV_{\alpha_j}(1)$.

Step 2. Calculate $SCV(\alpha_k, \mathbf{x}_{i_k})$ defined in (1.13) with α replaced by α_k and z by $\mathbf{x}_{i_k}$.

If $CV_{\alpha_k}(1) > SCV(\alpha_k, \mathbf{x}_{i_k})$, stop and model $\mathcal{M}_\alpha$ is selected. Otherwise, go to step 1 with α updated with α_k. Repeat the above procedures until no further variables can be removed. We call this selection procedure the separated cross-validation (SCV) method.

Remark If (1.10) holds, then by (1.8) and (1.9), we have

$$P\{CV_{\alpha_1}(1) > CV_{\alpha_2}(1)\} \to 0, \quad \text{for any } M_{\alpha_1} \text{ in Category 1 and } M_{\alpha_2} \text{ in Category 2.}$$

This implies that if $\alpha_0 \subset \alpha$, then after Step 1, we still have $\alpha_k \supseteq \alpha_0$ in probability and consequently by Theorem 1.1, the output of Step 2 will be α_k in probability. That is, we

Table 1.1: Penalties for some variable selection criteria with sample size n.

Method	Asymptotic Expansion	Penalty	Consistency
AIC	$\log(RSS^*) + 2d/n + constant$	$2d/n$	N
BIC	$\log(RSS^*) + \log(n)d/n + constant$	$\log(n)d/n$	Y
$CV(1)$	$RSS^* + \sigma^2(2d - \epsilon' P_\alpha \epsilon)/n$	$\sigma^2\{2 - \chi^2(1)\}/n$	N
$CV(m)$	$RSS^* + \sigma^2/(n-m)$	$\sigma^2/(n-m)$	Y
SCV	$RSS^* + \sigma^2(R_K + 4c_k)/(nh)$	$\sigma^2 R_K/(nh)$	Y

RSS^*, residual sum of squares under model $\mathcal{M}_{\alpha_0}$, $d = d_\alpha$.

successfully locate one extra variable contained in $\mathcal{M}_\alpha$. Theorem 1.1 also implies that if $\alpha = \alpha_0$, then no variable will get removed after Step 2 with probability tending to 1.

Theorem 1.1 *Suppose (1.7) (1.10) and* $(A1) - (A4)$ *in Appendix A hold, then*

$$SCV(\alpha, \mathbf{x}_k) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 + \frac{\sigma^2(R_K + 4c_k)}{nh} + \beta_k \Delta + o_p(\frac{1}{nh}), \tag{1.14}$$

where $R_K = \int K^2(v)dv$, c_k *is the Lebesgue measure of the support of* $\mathbf{x}_k$ *and* $\Delta \geq 0$ *with* $E\Delta = o(n^{-1}h^{-b})$ *for any* $b > 1$. *Therefore,*

1. *If* $k \in \alpha_0$ *and* $\alpha \cup k = \alpha_0$, *then* $\lim_{n\to\infty} Pr\{SCV(\alpha, \mathbf{x}_k) > CV_{\alpha_0}(1)\} \to 0$.

2. *If* $\alpha_0 \subseteq \alpha$, *and* $k \notin \alpha_0$, *then* $\lim_{n\to\infty} Pr\{SCV(\alpha, \mathbf{x}_k) < CV_\alpha(1)\} \to 0$.

The proof of the theorem is given in Appendix A. To get a feel of the newly proposed SCV method and some popular selection criteria, we list in Table 1.1 the asymptotic expansion form of each method and its penalty term, when one redundant variable is added into the true model $\mathcal{M}_0$. We can see that the 'penalty' imposed by those consistent methods are invariably larger than by inconsistent ones.

1.4 Simulations and Examples

We study the finite sample performance of several selection criteria, namely AIC, BIC, $CV(1)$, $CV(m)$ and SCV. In all the calculations below, the Epanechnikov kernel $K(\mu) = 0.75(1-\mu^2)_+$ is used. Since the function $g(.)$ in (1.12) is actually linear, the optimal bandwidth which minimizes the mean squared error is infinite([26]). Fortunately, the choice of bandwidth in subset selection is not as crucial as in smoothing regression, as long as the order of the bandwidth meets the requirement for consistency ([13, 89]). Therefore, it suffices to use the rule-of-thumb ([77], pp.45-7), thus bypassing the aforementioned problem.

To take into account of the variation of y, we propose to use the following scheme for the partially linear model $Y_i = \beta^\top X_i + g(z_i) + \varepsilon_i, \quad i = 1, \cdots, n$. First calculate the residual errors of a linear regression model of Y on X as

$$\hat{e}_i = y_i - n^{-1} X_i^\top \Big(n^{-1} \sum_{i=1}^n X_i X_i^\top\Big)^{-1} n^{-1} \sum_{i=1}^n X_i Y_i.$$

Then compute the conditional variance of $\hat{e}_i$ on z_i using the method in [32] as

$$\hat{\sigma}_e^2 = \frac{1}{n-2} \sum_{i=1}^{n-2} (0.809\hat{e}_i - 0.5\hat{e}_{i+1} - 0.309\hat{e}_{i+2})^2.$$

Then the bandwidth is chosen to be $h = \sigma_x \sigma_y^{-1} \hat{\sigma}_e / n^{0.2}$, which is parallel to the one proposed by Silverman ([77]) with coefficient adjusted.

Table 1.2: Simulation results for Example 1.2

Methods	$n = 40, \sigma = 3$			$n = 40, \sigma = 1$			$n = 60, \sigma = 1$			$n = 100, \sigma = 1$		
	(M)	(T)	(E)	(M)	(T)	(E)	(M)	(T)	(E)	(M)	(T)	(E)
AIC	.31	4.04	.20	.40	4.12	0	.41	4.15	0	.41	4.20	0
BIC	.49	4.60	.33	.72	4.69	0	.78	4.76	0	.90	4.90	0
$CV(1)$	.37	4.14	.17	.45	4.19	0	.45	4.18	0	.46	4.26	0
$CV(m)$	.30	4.79	.64	.82	4.79	0	.82	4.81	0	.86	4.84	0
SCV	.37	4.79	.69	.93	4.94	.01	.96	4.96	0	1	5	0

Example 1.2 In this example we simulated 100 data sets with sample size n from model

$$Y = X^\top \beta + \sigma\varepsilon, \quad \beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top,$$

where the components of X and ε are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with ρ= 0.5. This is a model used in [81]. The model error of the proposed procedures is compared to that of some other methods. The column labeled '(M)' in Table 1.2 is the frequency of correct model selection. The average number of zero coefficients is also recorded. The column labeled '(T)' are the average restricted only to the true zero coefficients and the column labeled '(E)' are for coefficients erroneously set to 0. From inspection of Table 1.2, we can see that SCV method performs the best.

Example 1.3 We consider the model in [73]

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i$$

where $i = 1, \cdots, 40$, e_i are $i.i.d.$ $N(0, 1)$, x_{ki} is the ith value of the kth prediction

Table 1.3: Simulation results for Example 1.3

True model	AIC	BIC	$CV(1)$	$CV(m)$	SCV
(2 0 0 4 0)	.588	.856	.484	.934	.882
(2 0 0 4 8)	.690	.866	.641	.947	.858
(2 9 0 4 8)	.996	.996	.801	.965	.968
(2 9 6 4 8)	1.000	1.000	.985	.948	.920

variable $x_k,\ x_{1i} \equiv 1$ and the values of $x_{ki},\ k = 2, \cdots, 5,\ i = 1, \cdots, 40$, are taken from the example in [31]. Here we only consider four different models with at least three nonzero β_k's and this is in favor of those methods with relatively lighter penalty, such as AIC, BIC and $CV(1)$. Frequencies out of 500 simulations that the true model is selected are recorded in Table 1.3.

Example 1.4 [Ground Ozone Level] Air pollution has serious impact on the health of plants and animals (including human beings) and reduces the visibility; see the report of WHO ([88]). Substances not naturally found in the air or at greater concentrations or in different locations from usual are referred to as 'pollutants'. The main pollutants include nitrogen dioxide (NO_2), Carbon dioxide (CO), sulphur dioxide (SO_2), respirable particulate (PM), ozone (O_3) and others. Pollutants can be classified as either primary or secondary. Primary pollutants are substances directly produced by a process, such as ash from a volcanic eruption or the carbon monoxide gas from a motor vehicle exhaust. Secondary pollutants are not emitted, such as ozone, which is produced from the photochemical oxidation of volatile organic compounds in the presence of sunlight and nitrogen oxides and other pollutants. Let N, S, P, T and H be the weekly average levels of NO_2, SO_2, PM, temperature and humidity respectively. To account for the interaction effect, we take on the interaction between any two of them, resulting in 15 covariates altogether.

To decide which of the 15 covariates significantly contribute to the average level of ozone, we use the pollution data collected in Hong Kong from 1994 to 1997. A linear model with all 15 variables shows that linear regression is enough. The selection process of SCV are put in Table 1.4. Begin with the full covariate set of 15 variables. Its subset α_k obtained in Step 1 is given immediately below, with corresponding $CV_{\alpha_k}(1)$ and $SCV(\alpha_k, x_k)$ value put in the neighbor columns labeled '$CV(1)$' and 'SCV' respectively. For example, among all size 3 subsets of $(H, N*T, S*H, P*T)$, $(H, S*H, P*T)$ has the smallest $CV(1) = 0.3513$ and the SCV value is 0.3110 for model

$$y = (H, S*H, P*T)*\beta + g(N*T) + \epsilon.$$

Focusing on the column labeled '$CV(1)$', we can see that the backward $CV(1)$ will pick up $H, N*S, N*T, S*T, S*H, P*T$ and $T*H$ (values in italic), although it makes no sense to take in the chemistry interaction factor $N*S$. Variables selected by SCV are weather conditions and their interactions with NO_2 and SO_2 (values in boldface), which is in line with the chemical claim that ozone is produced from chemical reactions between reactive organic gases (P) and oxides of nitrogen in the presence of sunlight.

Table 1.4: Variable selection procedure for Example 1.4

Variable candidates	$CV(1)$	SCV
$(N, S, P, T, H, N*S, N*P, N*T, N*H, S*P, S*T, S*H, P*T, P*H, T*H)$	0.2992	0.3161
$(S, P, T, H, N*S, N*P, N*T, N*H, S*P, S*T, S*H, P*T, P*H, T*H)$	0.2945	0.3059
$(S, P, H, N*S, N*P, N*T, N*H, S*P, S*T, S*H, P*T, P*H, T*H)$	0.2916	0.3106
$(P, H, N*S, N*P, N*T, N*H, S*P, S*T, S*H, P*T, P*H, T*H)$	0.2895	0.3013
$(P, H, N*S, N*P, N*T, N*H, S*T, S*H, P*T, P*H, T*H)$	0.2895	0.2943
$(P, H, N*S, N*T, N*H, S*T, S*H, P*T, P*H, T*H)$	0.2880	0.3183
$(P, H, N*S, N*T, S*T, S*H, P*T, P*H, T*H)$	0.2868	0.2895
$(H, N*S, N*T, S*T, S*H, P*T, P*H, T*H)$	0.2859	0.2864
$(H, N*S, N*T, S*T, S*H, P*T, T*H)$	*0.2842*	0.2945
$(H, N*S, N*T, S*H, P*T, T*H)$	0.2859	0.2963
$(H, N*S, N*T, S*H, P*T)$	0.2941	0.3099
$(H, N*T, S*H, P*T)$	**0.2939**	**0.3110**
$(H, S*H, P*T)$	**0.3513**	0.3517

Chapter 2

Subset Selection for Single-Index Models

2.1 Introduction

As a semi-parametric approach attending to tackle data in high dimensions, the single-index model (SIM) is widely used in applied quantitative sciences, such as econometrics and statistics. Suppose Y is a response variable and $X = (\mathbf{x}_1, \cdots, \mathbf{x}_p)^\top$ are covariates. The single-index model is written as

$$Y = g(X^\top \theta^0) + \varepsilon, \tag{2.1}$$

where $E(\varepsilon|X) = 0$ almost surely, g is an unknown link function and θ^0 is an unknown unit parameter vector (single-index) with its first nonzero component positive for identification purposes. Many widely parametric models have this form; examples include linear regression, binary logit and probit and Tobit models. These models assume that g is known; when g is unknown, SIM is more flexible than a parametric model while avoiding the loss of precision that occurs in fully nonparametric estimation with a multidimensional X.

Recent papers ([34, 36, 41, 42, 44, 45, 65, 92]) have considered the estimation of the parametric index and the nonparametric link function with focus on the root-n consistency of the former; efficiency issues have also been studied. Amongst them, the most popular ones are the sliced inverse regression method ([48]), the semi-parametric least squares estimator ([34, 44]) and the minimum average conditional variance estimator ([92]). If X are continuous, then the computation difficulty can be greatly reduced through the use of average derivative estimator (ADE, [36]), which relies on the fact that $E[\partial g(X^\top \theta)/\partial X] \propto \theta$. However, because the high dimensional kernel estimation method is used, the estimation still suffers from the so called "curse of dimensionality". [42] adopted the same idea as ADE and came up with a dynamic procedure to adapt to the structure of the model by lowering the dimension of the kernel smoothing. On the other hand, to tackle the situation when $E[\partial g(X^\top \theta)/\partial X] = 0$, [91] proposed an outer product of gradients method which is based on the fact that θ is the eigenvector corresponding to the greatest eigenvalue of $E[\partial g(X^\top \theta)\partial^\top g(X^\top \theta)]$.

All the studies mentioned above assume that all regressors X contain useful information for predicting the response variable. If irrelevant regressors are included, which is very likely in high dimensional environments ([59]), the precision of parameter estimation as well as the accuracy of forecasting will suffer ([2]). Therefore, it is necessary to remove irrelevant variables from SIM. Using sliced inverse regression method (SIR), [59] considered this issue when the error term ϵ is normally distributed and X are continuous and elliptically symmetric. However, in practice it is common that some covariates are asymmetric or discrete. In this case, SIR fails to obtain a useful estimator of the single-index parameter and the method of [59] is thus inapplicable.

Cross-validation method and its equivalent have long been used in model identification and subset selection ([58]). We mentioned in Chapter 1 that under nonparametric settings, the leave-one-out CV method is consistent. In fact, the same result holds for 'leave-m-out' CV (CV(m)), the proof of which is given in Appendix B.

Semi-parametric models are different again. I will prove that CV(1) again fails to select variables in SIM but CV(m) does not, provided that $m/n \to c \in [2/3, 1)$, different from the requirements on m in linear regression models. Thus no more than 1/3 of the samples can be used for model estimation and this is usually not enough to estimate the model well, resulting in inferior efficiency in variable selection. Furthermore, CV(m) is computationally prohibitive. To overcome these disadvantages, we shall propose a new variable selection method called separated cross-validation method (SCV).

2.2 Optimal Model and Parameter Estimation

We use notation similar to that in [73]. Let $\mathcal{S}$ denote all nonempty subsets of $\{1, \cdots, p\}$. For any $\alpha \in \mathcal{S}$, let d_α be the cardinality of α, θ_α and X_α be two $d_\alpha \times 1$ column vectors, which containing the components of θ or X indexed by the integers in α respectively. Let

θ denote the vector which minimizes $E[Y - E(Y|X_\alpha^\top \theta)]^2$. The corresponding single-index model

$$Y = g_\alpha(X_\alpha^\top \theta) + \epsilon_\alpha, \quad \epsilon_\alpha = Y - E(Y|X_\alpha^\top \theta) = Y - g_\alpha(X_\alpha^\top \theta) \tag{2.2}$$

is denoted by $\mathcal{M}_\alpha$. If we know whether or not each component of θ^0 is 0, then models $\mathcal{M}_\alpha$ can be classified into two categories. In one category, at least one covariate with a nonzero coefficient in (2.1) is missing in X_α. In the other category, X_α contains all covariates with nonzero coefficients. The true model denoted by $\mathcal{M}_{\alpha_0}$, is defined as the model in the second category with the smallest number d_0 of covariates.

Suppose $\{(X_i, Y_i), i = 1, \cdots, n\}$ is a random sample from model (2.1). Consider model $\mathcal{M}_\alpha$ with $\alpha_0 \subseteq \alpha$. To guarantee the consistency of estimation, we assume throughout the paper that $X_\alpha^\top \theta_\alpha$ has an almost everywhere positive density function for any $\alpha \supseteq \alpha_0$ and θ in a small neighborhood of θ_α^0, a column vector containing the components of θ^0 indexed by the integers in α; see [41] for more discussion. The popular method proposed by [34] estimates the model as follows. Suppose $A \subseteq R^p$ is a compact convex set such that the density function of $X^\top \theta$ is uniformly bounded away from zero on $\{\theta^\top x : x \in A\}$ for any θ near θ^0. For any given $b > 0$ and $h > 0$, let $A^{bh} = \{x \in R^p : \|x - x_0\| \leq bh$ for some $x_0 \in A\}$. The introduction of A and A^{bh} is for technical purposes; see [34] for more details. Let $g_\alpha(u|\theta) = E(Y|X_\alpha^\top \theta = u^\top \theta)$. Its leave-one-out estimate is given by

$$\hat{g}_\alpha^{\setminus i}(u|\theta) = \frac{\sum_{j \neq i} K_h(X_{j,\alpha}^\top \theta - u^\top \theta) Y_j}{\sum_{j \neq i} K_h(X_{j,\alpha}^\top \theta - u^\top \theta)}, \tag{2.3}$$

where h is a bandwidth, K is an univariate density function with support $[-b, b]$ and $K_h(.) = h^{-1}K(./h)$. Since $g_\alpha(u|\theta_\alpha^0) \equiv g(u^\top \theta_\alpha^0)$, the index parameter under model $\mathcal{M}_\alpha$ is estimated by minimizing

$$HCV_\alpha(\theta, h) \triangleq {\sum_i}' \{Y_i - \hat{g}_\alpha^{\setminus i}(X_{i,\alpha}|\theta)\}^2, \tag{2.4}$$

with respect to θ and $h > 0$ subjected to $\|\theta\| = 1$, where ${\sum_i}'$ denotes summation over indices i such that $X_i \in A$. We assume that all $X_i \in A^{bh}$. Otherwise one can always completely ignore those data outside of A^{bh}. To make the notations neat, let $X_i \in A$ if $1 \le i \le n'$ and $X_i \notin A$ if $i > n'$, which implies that $n - n' = O(nh)$. This estimator has very good asymptotic properties. It needs no under-smoothing for the estimator of θ to achieve root-n consistency. However, it is not easy to solve the above minimization problem, even when $d_\alpha = 2$, let alone even higher dimensions.

Based on local linear approximation, [92] estimated θ_α^0 by

$$\hat{\theta} = \arg \min_{\theta:\|\theta\|=1} \sum_{j=1}^n \sum_{i=1}^n (Y_i - a_j - d_j \theta^\top X_{ij,\alpha})^2 w_{ij},$$

where $X_{ij,\alpha} = X_{i,\alpha} - X_{j,\alpha}$ and w_{ij} is a weight depending on the distance between $X_{i,\alpha}$ and $X_{j,\alpha}$. The corresponding algorithm takes the following form; with an initial value θ, calculate

$$\begin{pmatrix} a_j^\theta \\ d_j^\theta \end{pmatrix} = \left[\sum_i K_h(X_{ij,\alpha}^\top \theta) \begin{pmatrix} 1 \\ X_{ij,\alpha}^\top \theta \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij,\alpha}^\top \theta \end{pmatrix}^\top \right]^+ \sum_i K_h(\theta^\top X_{ij,\alpha}) \begin{pmatrix} 1 \\ X_{ij,\alpha}^\top \theta \end{pmatrix} Y_i \tag{2.5}$$

and then calculate

$$\theta = \Big[\sum_{i,j} K_h(X_{ij,\alpha}^\top \theta)(d_j^\theta)^2 X_{ij,\alpha} X_{ij,\alpha}^\top \Big]^+ \sum_{i,j} K_h(X_{ij,\alpha}^\top \theta) d_j^\theta X_{ij,\alpha}(Y_i - a_j^\theta), \ \theta = sign(\theta_1)\frac{\theta}{\|\theta\|}, \tag{2.6}$$

where $[.]^+$ denotes the Moore-Penrose inverse of the matrix in the brackets. Repeat (2.5) and (2.6) until the iteration process converges, to what we call the minimum average variance estimator (MAVE).

[91] proved that the MAVE estimator is root-n consistent and has the same asymptotic distribution as the estimator of [34], referred to as HHI herein.

2.3 Cross-validation Subset Selection

In the cross-validation method, the data are split into two sets, training set s^c and the test set s. The training set is used to estimate all candidate models and the model that best predicts the test set is the preferred model.

2.3.1 CV(m) Based on HHI Method

The leave-m-out HHI estimator of $g_\alpha(\mu|\theta)$ is given by

$$\hat{g}_\alpha^{\setminus s}(u|\theta) = \sum_{j \notin s} Y_j K_h(u - X_{j,\alpha}^\top \theta) / \sum_{j \notin s} K_h(u - X_{j,\alpha}^\top \theta),$$

where s is a subset of $\{1, \cdots, n'\}$ and $\#s = m$. We then estimate θ_α^0 by minimizing

$$HCV_\alpha^m(\theta, h) \triangleq \frac{1}{n_v \binom{n}{n_v}} \sum_s \sum_{i \in s} \{Y_i - \hat{g}_\alpha^{\setminus s}(X_{i,\alpha}^\top \theta|\theta)\}^2, \tag{2.7}$$

where summation $\sum\limits_s$ runs over all possible size m subsets of $\{1, \cdots, n'\}$. Let $HCV_\alpha^m = \min\limits_{\theta,h} HCV_\alpha^m(\theta, h)$. The following theorem shows that HCV_α^m can not be used for subset selection.

Theorem 2.1 *If* $(A1) - (A4)$ *in Appendix B hold and* $m/n \to c \in [0, 1)$, *then for any* $\alpha \supset \alpha_0$, $Pr\{HCV_\alpha^m < HCV_{\alpha_0}^m\} \to 1$, *as* $n \to \infty$.

2.3.2 CV(m) Based on MAVE Method

Note that in (2.2), $\theta = \theta^0_\alpha$ for any $\alpha \supset \alpha_0$. For such α and any $s \subset \{1, \cdots, n'\}$ with $\#s = m$, we first estimate θ^0_α by $\hat{\theta}^{\backslash s}_\alpha$, the MAVE estimator of the index vector θ in model (2.2) from $\{(X_j, Y_j) : 1 \leq j \leq n,\ j \notin s\}$. The link function is then estimated by the local linear smoother

$$\hat{g}^{\backslash s}_\alpha(u|\hat{\theta}^{\backslash s}_\alpha) = \sum_{j\notin s} M_{\alpha,h}((X_{j,\alpha} - u)^\top \hat{\theta}^{\backslash s}_\alpha) Y_j \Big/ \sum_{j\notin s} M_{\alpha,h}((X_{j,\alpha} - u)^\top \hat{\theta}^{\backslash s}_\alpha), \tag{2.8}$$

where

$$\begin{aligned} M_{\alpha,h}((X_{j,\alpha} - u)^\top \hat{\theta}^{\backslash s}_\alpha) \quad = \quad & S^{\backslash s}_{\alpha,2}(u|\hat{\theta}^{\backslash s}_\alpha) K_h \left\{(X_{j,\alpha} - u)^\top \hat{\theta}^{\backslash s}_\alpha\right\} \\ & - S^{\backslash s}_{\alpha,1}(u|\hat{\theta}^{\backslash s}_\alpha) \left\{(X_{j,\alpha} - u)^\top \hat{\theta}^{\backslash s}_\alpha / h\right\} K_h \left\{(X_{j,\alpha} - u)^\top \hat{\theta}^{\backslash s}_\alpha\right\} \end{aligned}$$

with $S^{\backslash s}_{\alpha,k}(u|\theta) = \sum_{j\notin s} K_h \left\{(X_{j,\alpha} - u)^\top \theta\right\} \left\{(X_{j,\alpha} - u)^\top \theta / h\right\}^k$, $k = 0, 1, 2$. We define the leave-m-out cross-validation estimate of prediction error as

$$CV_\alpha(m) \overset{\triangle}{=} m^{-1} \binom{n'}{m}^{-1} \sum_s{}' \sum_{i\in s} \{Y_i - \hat{g}^{\backslash s}_\alpha(X_{i,\alpha}|\hat{\theta}^{\backslash s}_\alpha)\}^2, \tag{2.9}$$

where $\sum_s{}'$ indicates summation over all possible size m subsets of $\{1, \cdots, n'\}$. Later, we will use $\sum_{i,s}{}'$ to denote $\sum_s{}' \sum_{i\in s}$. The model $\mathcal{M}_\alpha$ with the smallest value of $CV_\alpha(m)$ is the selected model.

Theorem 2.2 *Suppose(A1)-(A5) in Appendix B hold. If $m \to \infty$, $m/n \to c \in [0, 1)$ and $h \propto n^{-1/5}$, then for any $\alpha \supset \alpha_0$ with $\delta_d := d_\alpha - d_0$, we have*

$$\lim_{n\to\infty} Pr\{CV_\alpha(m) > CV_{\alpha_0}(m)\} = Pr\{\chi^2(\delta_d) > \frac{(2 - 3c)\delta_d}{1 - c}\}.$$

By Theorem 2.2, for CV(m) to be consistent, i.e. $\lim_{n\to\infty} \Pr\{CV_\alpha(m) > CV_{\alpha_0}(m)\} = 1$, it is required that $2-3c \le 0$, or $1 > c \ge 2/3$. Although we have no conclusion in the case $c = 1$, our conjecture is that the consistency does not hold, since $\hat{\theta}_\alpha^{\backslash s}$ is no longer root-n consistent as $n_c := n - m = o(n)$, i.e. the size of learning set is much smaller than n.

The way CV(m) splits the data is acceptable for linear regression models, whose parameter can be estimated well with a small sample. However, the size of the training set used by CV(m) is usually too small for nonparametric smoothing methods. Another disadvantage of CV(m) is the heavy computational burden since there are $\binom{n'}{m}$ possible splitting combinations. To tackle this problem, Monte Carlo CV(m) randomly draws, with or without replacement, a collection $\mathcal{R}$ of subsets of $\{1, \cdots, n'\}$ of size m, and selects a model that minimizes

$$CV_\alpha^{\mathrm{mc}}(m) \triangleq \sum_{s\in\mathcal{R}} \sum_{i\in s} \{Y_i - \hat{g}_\alpha^{\backslash s}(X_{i,\alpha}|\hat{\theta}_\alpha^{\backslash s})\}^2.$$

In linear regression models, the performance of this method has been proved to be similar to that of CV(m); see [73, 94]. The Monte Carlo CV(m) is thus used in the simulation study instead of CV(m).

Although Theorem 2.2 is proved for MAVE estimator, the same results hold for other single-index model estimation methods, providing that the estimator has a similar stochastic expansion to that given in (B.1). Examples are the estimator by [34], albeit computationally intensive, and the average derivative estimator by [36]. The method of [42] might also be used as [91] proved that an alternative version has a similar expansion.

2.4 Subset Selection by Separation

Starting with the full covariate set $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$, we need to check whether or not a certain covariate, $\mathbf{x}_k$ say, contributes to the response variable Y. For this purpose, we introduce the following model

$$Y = g(X_\alpha^\top \theta, \mathbf{x}_k) + \epsilon, \ \alpha \cup k = \{1, \cdots, p\}. \tag{2.10}$$

Compared with model (2.1), where the contribution of $\mathbf{x}_k$ is mixed up with that of the other covariates through a linear combination, $\mathbf{x}_k$ in model (2.10) is 'separated' and its contribution can be assessed more accurately. Another reason for the introduction of model (2.10) is the different behavior of cross-validation method in parametric models and nonparametric models. As the relationship between Y and $\mathbf{x}_k$ is 'nonparametric' in (2.10), simple CV(1) can tell whether or not $\mathbf{x}_k$ contributes to Y as proved in [13, 89].

The parameter θ in model (2.10) can be estimated by the first d_α entries of the MAVE estimator of the index vector in SIM $Y = g(X_{\alpha\cup k}^\top \theta) + e$. For any fixed θ, define $g_{\alpha,k}(u, v|\theta) = E(Y|X_\alpha^\top \theta = u^\top \theta, \mathbf{x}_k = v)$. Its leave-one-out local linear estimator is the first component of

$$\left\{ \sum_{j\neq i} K_{h_1,j}^{\alpha,\theta}(u,v) \begin{pmatrix} 1 \\ \theta^\top (X_{j,\alpha} - u) \\ X_{j,k} - v \end{pmatrix} \begin{pmatrix} 1 \\ \theta^\top (X_{j,\alpha} - u) \\ X_{j,k} - v \end{pmatrix}^\top \right\}^{-1} \sum_{j\neq i} K_{h_1,j}^{\alpha,\theta}(u,v) \begin{pmatrix} 1 \\ \theta^\top (X_{j,\alpha} - u) \\ X_{j,k} - v \end{pmatrix}^\top Y_j, \tag{2.11}$$

where $K_{h_1,j}^{\alpha,\theta}(u,v) = K_{h_1}(\theta^\top X_{j,\alpha} - u) H_{h_1}(\mathbf{x}_{j,k} - v)$ is a two-dimensional product kernel, h_1 is a bandwidth and $H = K$ for $\mathbf{x}_k$ continuous and $H_h(v) = I(v = 0)$ for $\mathbf{x}_k$ discrete. To make notations more neat, let $\hat{g}_{\alpha_1,k}^{\setminus i}(X_i|\hat{\theta}_{\alpha_1}^{\setminus i})$ and $\hat{g}_{\alpha_1}^{\setminus i}(X_i|\hat{\theta}_{\alpha_1}^{\setminus i})$ denote $\hat{g}_{\alpha_1,k}^{\setminus i}(X_{i,\alpha_1}, \mathbf{x}_{i,k}|\hat{\theta}_{\alpha_1}^{\setminus i})$

and $\hat{g}^{\backslash i}_{\alpha_1}(X_{i,\alpha_1}|\hat{\theta}^{\backslash i}_{\alpha_1})$ respectively. We propose the following algorithm for variable selection. Start with an initial covariate set α satisfying $\alpha_0 \subseteq \alpha$.

Step 1. Calculate $\hat{\theta}$, the MAVE estimator of θ in model $Y = g(X_\alpha^\top \theta) + \epsilon$ from all data points. Find the entry of $\hat{\theta}$ with the smallest absolute value and its corresponding index in α, k say. Set $\alpha_1 = \alpha \setminus \{k\}$.

Step 2. Denote by $\hat{\theta}^{\backslash i}_\alpha$ the MAVE estimator of θ in $Y = g(X_{\alpha_1 \cup k}^\top \theta) + \epsilon$ based on $\{(X_j, Y_j)\}_{j \neq i}$. Eliminate the last entry and denote the rest by $\hat{\theta}^{\backslash i}_{\alpha_1}$.

Step 3. Calculate $\hat{g}^{\backslash i}_{\alpha_1,k}(X_i|\hat{\theta}^{\backslash i}_{\alpha_1})$ as defined in (2.11) and $\hat{g}^{\backslash i}_{\alpha_1}(X_i|\hat{\theta}^{\backslash i}_{\alpha_1})$ as defined in (2.5), with θ, and α replaced by $\hat{\theta}^{\backslash i}_{\alpha_1}$ and α_1 respectively. Let

$$CV_{\alpha_1,k} = \frac{1}{n'}\sum_i{}' \{Y_i - \hat{g}^{\backslash i}_{\alpha_1,k}(X_i|\hat{\theta}^{\backslash i}_{\alpha_1})\}^2, \; CV_{\alpha_1} = \frac{1}{n'}\sum_i{}' \{Y_i - \hat{g}^{\backslash i}_{\alpha_1}(X_i|\hat{\theta}^{\backslash i}_{\alpha_1})\}^2,$$

where $\sum_i{}'$ is defined in (2.4). If $CV_{\alpha_1,k} < CV_{\alpha_1}$, stop and select model α. Otherwise go to Step 1 with α replaced by α_1.

Repeat the above procedure until no further variable can be removed. We call this procedure the separated cross-validation method (SCV).

Step 1 is employed to simplify the calculations. As θ^0 can be estimated with root-n consistency in SIM, if $\alpha \supset \alpha_0$, then $\hat{\theta}_k = O_p(n^{-1/2})$; if $\mathbf{x}_k$ is necessary, $\hat{\theta}_k = \theta^0_k + O_p(n^{-1/2})$, which is bounded away from 0 in probability. Therefore, if α is too large, then with probability tending to 1, only its redundant variables will be considered for removal by Step 1. Computations in Steps 2 and 3 can also be simplified by replacing $\hat{\theta}^{\backslash i}_\alpha$ and $\hat{\theta}^{\backslash i}_{\alpha_1}$ with $\hat{\theta}_\alpha$ and $\hat{\theta}_{\alpha_1}$ respectively. Step 2 estimates the parameters in model (2.10) assuming that $\mathbf{x}_k$ should be removed. Step 3 compares the cross-validation values for model (2.10) and (2.2) to check the importance of $\mathbf{x}_k$; see [13].

As shown in [34, 91], the same bandwidth can be used to estimate the link function as well as the index parameter. To implement (2.11), theoretical justification requires different bandwidths used for the estimation of model (2.10), depending on the type of $\mathbf{x}_k$: $h_1 \propto n^{-1/6}$ for $\mathbf{x}_k$ continuous and $h_1 = h \propto n^{-1/5}$ for $\mathbf{x}_k$ discrete, where h is the bandwidth used when calculating CV_{α_1}. Many available bandwidth selection methods, such as the cross-validation, the generalized cross-validation, and the rule-of-thumb can be used to choose the bandwidths; see [26, 77] for more details. More is to be said about this in Section 5 below. We have the following consistency property for SCV method.

Theorem 2.3 *Suppose assumptions (A1)-(A7) in Appendix B hold and that the bandwidth satisfies the requirement mentioned above.*

1. *If* $\alpha \cup k = \alpha_0$, *then* $\lim_{n\to\infty} Pr\{CV_{\alpha,k} > CV_\alpha\} \to 0$.

2. *If* $\alpha_0 \subseteq \alpha$ *and* $k \notin \alpha_0$, *then* $\lim_{n\to\infty} Pr\{CV_{\alpha,k} < CV_\alpha\} \to 0$.

2.5 Simulation Study

We compare CV(1), CV(m) and SCV by simulations. Since the asymptotic distribution of $\hat{\theta}$ can be used for variable selection, we also include it in the comparison study. The distributional result is that

$$n^{1/2}(\hat{\theta} - \theta^0) \rightarrow N(0, W_0^+ W_1 W_0^+), \tag{2.12}$$

in distribution as $n \to \infty$, where $W_0 = E[\{X - E(X|X^\top\theta^0)\}\{X - E(X|X^\top\theta^0)\}^\top g'(X^\top\theta^0)^2]$, $W_1 = E[\{X - E(X|X^\top\theta^0)\}\{X - E(X|X^\top\theta^0)\}^\top g'(X^\top\theta^0)^2\varepsilon^2]$ and W_0^+ denotes the Moore-Penrose inverse. The matrices W_0 and W_1 can be estimated by kernel smoothing as $\hat{W}_0 = n^{-1}\sum_{i=1}^n (X_i - \hat{\mu}_i)(X_i - \hat{\mu}_i)^\top \hat{d}_i^2$ and $\hat{W}_1 = n^{-1}\sum_{i=1}^n (X_i - \hat{\mu}_i)(X_i - \hat{\mu}_i)^\top \hat{d}_i^2 (Y_i - \hat{a}_i)^2$, where $\hat{\mu}_i = \sum_{j=1}^n K_h(X_{ij}^\top\hat{\theta})\ X_j / \sum_{j=1}^n K_h(X_{ij}^\top\hat{\theta})$ with $\hat{a}_i$ and $\hat{d}_i$ given by

$$\begin{pmatrix}\hat{a}_i\\ \hat{d}_i\end{pmatrix} = \Big\{\sum_{j=1}^n K_h(X_{ij}^\top\hat{\theta})\begin{pmatrix}1\\ X_{ij}^\top\hat{\theta}\end{pmatrix}\begin{pmatrix}1\\ X_{ij}^\top\hat{\theta}\end{pmatrix}^\top\Big\}^{-1}\sum_{j=1}^n K_h(X_{ij}^\top\hat{\theta})\begin{pmatrix}1\\ X_{ij}^\top\hat{\theta}\end{pmatrix}Y_j.$$

Based on (2.12), a variable $\mathbf{x}_k$ is selected if $|\hat{\theta}_k| > 1.96(c_{kk}/n)^{1/2}$, where c_{kk} is the (k,k) entry of $\hat{W}_0^+\hat{W}_1\hat{W}_0^+$.

In calculations below, the Gaussian kernel is used, since we find empirically it performs better in estimating the index parameter; see also [72]. After (X_i, y_i) are standardized, the bandwidths are calculated by the rule-of-thumb ([77], pp. 45-7) as follows. In (2.5), $h = 1.06 s_{\theta^\top X_\alpha} n^{-1/5}$, where $s_{\theta^\top X_\alpha}$ is the sample standard deviation of $\theta^\top X_{i,\alpha}$. In (2.11), $h_1 = 1.06\max(s_{\theta^\top X_\alpha}, 1)n^{-1/6}$ for $\mathbf{x}_k$ continuous and $h_1 = h$ for $\mathbf{x}_k$ discrete.

Example 2.4 We draw random samples with size $n = 50, 100$ and 200 respectively from a logistic regression model

$$Y \sim \mathrm{Ber}\{l(X^\top\beta)\}, \qquad l(\mu) = \exp(\mu)/\{1 + \exp(\mu)\},$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$. Two designs were used for $X = (\mathbf{x}_1, \cdots, \mathbf{x}_8)^\top$. In Design A, $(\mathbf{x}_1, \cdots, \mathbf{x}_6)^\top \sim N(0, \Sigma_6)$, where $\Sigma_p = (0.5^{|i-j|})_{1\le i\le j\le p}$, and $\mathbf{x}_7$, $\mathbf{x}_8$, are independent Ber(0.5), independent of $(\mathbf{x}_1, \cdots, \mathbf{x}_6)^\top$. In Design B, $\mathbf{x}_{(2k)} = 2I(\mathbf{z}_{(2k)} > 0) - 1$ and $\mathbf{x}_{(2k-1)} = \mathbf{z}_{(2k-1)}$, for $k = 1, 2, 3, 4$, where $Z = (\mathbf{z}_1, \cdots, \mathbf{z}_8) \sim N(0, \Sigma_8)$. Design A was investigated by [28]. A single-index model is fitted to the data and the variable

Table 2.1: Frequency of correct model selection for Example 2.4

Design	n	CV(1)	CV($0.25n$)	CV($0.5n$)	CV($0.75n$)	SCV	ASD
	50	0	0	0.18	0.38	0.41	0.27
(A)	100	0.29	0.46	0.58	0.63	0.66	0.44
	200	0.23	0.47	0.85	0.68	0.90	0.72
	50	0.3	0.37	0.46	0.46	0.51	0.32
(B)	100	0.37	0.43	0.69	0.77	0.81	0.65
	200	0.67	0.71	0.80	0.87	0.91	0.75

selection methods are applied. The relative frequencies of correct subset selection among 100 replications are reported in Table 2.5. We can see that SCV outperforms all the other methods. Its efficiency is even comparable with the results of [28], where the model is known up to unknown parameters. Also, the table shows that the CV(m) usually has better performance if the data is split in the way according to Theorem 2.2.

Example 2.5 The Tobit model is an econometric model in which the dependent variable is censored. In the original model in [83], for example, the dependent variable was expenditure on consumer durables, and the censoring occurs as values below zero are not observed, i.e.

$$Y = (\beta^\top X + 0.5\varepsilon) I(\beta^\top X + 0.5\varepsilon > 0), \tag{2.13}$$

where $I(.)$ is an indicator function; see also [61]. We also consider two designs: (A) $X = (\mathbf{x}_1, \cdots, \mathbf{x}_{20})^\top \sim N(0, I_{20})$; (B) $\mathbf{x}_{(2k)} = 2I(\mathbf{z}_{(2k)} > 0) - 1$ and $\mathbf{x}_{(2k-1)} = \mathbf{z}_{(2k-1)}$, for

Table 2.2: Frequency of correct model selection for Example 2.5

Design	l	n	CV(1)	CV($0.5n$)	CV($0.75n$)	SCV	ASD
(A)	5	50	0.08	0.36	0.02	0.84	0.08
	10	50	0.17	0.49	0.14	0.60	0.14
	5	100	0.32	0.82	0.78	0.99	0.26
	10	100	0.56	0.90	0.93	1.00	0.33
(B)	5	50	0.12	0.38	0.0	0.85	0.03
	10	50	0.14	0.32	0.0	0.59	0.17
	5	100	0.42	0.92	0.93	0.97	0.10
	10	100	0.55	0.92	0.90	0.99	0.37

$k = 1, \cdots, 10$, where $Z \sim N(0, \Sigma_{20})$. The error term $\varepsilon \sim N(0,1)$ is independent of X and $\beta = (1, 1, \cdots, 1, 0, \cdots, 0)^\top$ with first l elements 1 and others 0.

Here we have more covariates than in Example 1. As we mentioned at the beginning of Section 4, having a large number of covariates will compromise the efficiency of CV(m) and this is clearly reflected in Table 2.5, where the relative frequencies of correct subset selection among 100 simulations are recorded. We can see that CV($0.5n$) outperforms CV($0.75n$), suggesting that for small to medium sample size, the way of splitting the data suggested by Theorem 2.2 is not applicable due to the nature of nonparametric smoothing. In contrast, the SCV method is rather robust and performs better.

We also found from simulations not reported here that the choice of bandwidth is not as crucial in subset selection as in nonparametric regression. This phenomenon was also observed in [13]. As mentioned in Sections 3 and 4, other ways of estimating single-index models can also be used in CV(m) or SCV, but some can be very time consuming.

2.6 Applications to Two Real Data Sets

Example 2.6 (The Swiss banknotes data) The data contain 6 explanatory variables which are certain measurements of Swiss banknotes, called Length, Left, Right, Bottom, Top and Diagonal, and denoted by $\mathbf{x}_1, \cdots, \mathbf{x}_6$ respectively. The response variable Y is coded as 0 or 1, indicating whether a banknote is genuine or not. There are 200 banknotes, with the first 100 banknotes genuine and the others counterfeit.

The fitted values from single-index models using all variables, $(\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6)$, or variables selected by SCV are plotted in Fig. 2.1. The index parameters are estimated respectively as $\theta_{\mathrm{ALL}} = (-.1597, .4638, -.1549, .5699, .2922, -.5703)^\top$ and $\theta_{\mathrm{S}} = (.8006, .3011, -.5181)^\top$. Both models fit the data very well. To compare their prediction capabilities, we split the data randomly into a training set comprising 50 counterfeit banknotes and 50 genuine banknotes, and a test set containing the rest. We estimate the model with the training set, apply the estimated model to the test set and calculate the number of misspecifications. With different covariate sets, the average numbers of misspecifications based on 10000 replications of this random splitting are given in Table 2.6. A single-index model with variables selected by the principle component analysis is also compared; see [35]. Apparently SCV delivers the best results.

Table 2.3: The Swiss banknotes data: average numbers of misspecifications

Method	selected variables	Ave. No. of misspecifications
All variables	$\mathbf{x}_1, \cdots, \mathbf{x}_6$	0.5787
Cross-validation	$\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$	0.6223
SCV	$\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$	0.5100
Principle Component Anal.	$\mathbf{x}_5, \mathbf{x}_6$	0.5411

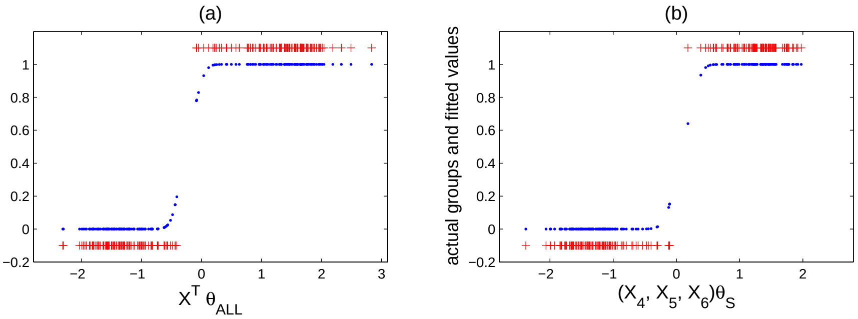

Figure 2.1: *'+': observations; '.': fitted values. (a) based on all covariates, (b) based on the selected variables. The observed Y are rescaled for easy visualization.*

Example 2.7 (The ozone concentration data) We study the relationship between ozone concentration level Y and radiation level R, temperature T, and wind speed W. 111 observations were taken daily from May to September 1973 in New York. Taking into account the interaction effect between any two covariates, we have totally nine covariates $X = (\mathbf{x}_1, \cdots, \mathbf{x}_9)^\top = (R, T, W, R^2, R * T, R * W, T^2, T * W, W^2)^\top$. After standardizing Y and $\mathbf{x}_k, k = 1, \cdots, 9$, we apply SCV to the data, thereby selecting variables $\mathbf{x}_3$, $\mathbf{x}_6$ and $\mathbf{x}_8$ with estimated index parameter $\theta_c = (.8486, -.0992, -.5196)^\top$. Single-index models with X or (R, T, W) as predictors are also investigated and the estimated index parameters are $\theta_b = (.2147, .1544, -.7541, -.1245, -.0029, -.0607, -.2292, .5183, .1448)^\top$, and $\theta_a =$

Table 2.4: Ozone concentration data: average prediction errors

Method	selected variables	Average prediction errors
All original variables	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	0.3643
All extended variables	$\mathbf{x}_1, \cdots, \mathbf{x}_9$	0.3621
SCV	$\mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_8$	0.3403

$(.3443, .7051, -.6199)^\top$ respectively. The fitted values from the three single-index models are plotted in Fig. 2.2. To compare the prediction capabilities of single-index models with different covariates, we again split the data randomly into two sets, this time with the training set comprising 56 observations and the test set containing the remaining 55 observations. The prediction errors are defined as the averaged sum of squared residuals. The results in Table 2.7 are based on 10000 replications of such random splitting.

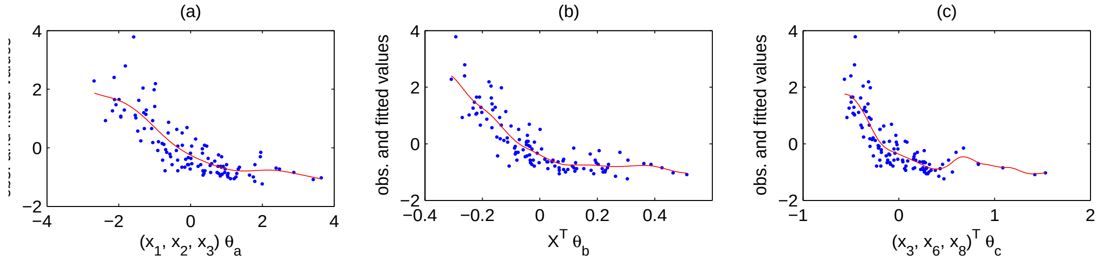

Figure 2.2: *'.': observations; '–': fitted values. (a) based on the original covariates* (R, T, W)*, (b) based on the extended variables, (c) based on the selected variables.*

Chapter 3

Conditional Heteroscedasticity Modeling in Finance

3.1 Introduction

In the 1970's, the autoregressive moving average processes (ARMA) was the focus of the research on time series modeling. Based on the conditional expectations, ARMA is easy to implement, with any temporal dependencies in the higher order moments treated as a nuisance. However, the three major drawbacks of ARMA models, namely linear setup, priori constraints on the parameters and conditional homoscedasticity, restrict the type of dynamics to be approximated and make this approach inadequate for structural interpretations. Among the fields of applications where standard ARMA fit is poor are financial and monetary problems. The financial time series features various forms of nonlinear dynamics, the crucial one being the strong dependence of the instantaneous variability (volatility or conditional heteroscedasticity) of the series on its own past, called the 'volatility clustering' phenomena. In simple words, that is, '...large changes tend to be followed by large changes of either sign, and small changes tend to be followed by small changes...'([52]). The panels on the left hand side of Fig. 3.6 show the daily returns (differenced in the logs of the daily closing price) of three stocks, with tickers IBM, BP and GM. Immediately evident is the existence of different regions where the daily returns and thus local volatility are relatively more or less extreme.

Understanding the exact nature of this temporal dependence in volatility is crucial for many issues in macroeconomics and finance, such as option pricing, the term structure of interest rates and risk management. For example, volatility is closely related to Value at Risk (VaR), the maximum loss over a given time horizon at a given confidence level α. In fact, VaR with confidence level α can be estimated by $\Phi^{-1}(\alpha)\hat{h}_t^{1/2}$, where $\Phi^{-1}(\alpha)$ is the α quantile of standard normal distribution and $\hat{h}_t$ is the estimated volatility.

The structure of this chapter is as follows. First, some well established parametric and nonparametric modeling of conditional heteroscedasticity, such as the ARCH, GARCH and stochastic volatility models, are briefly described in Sections 2 and 3. Section 4 is an empirical study of the nonparametric volatility model and GARCH(1,1) using an extensively investigated real data set. In Section 5, a new model called the Monotone constrained varying coefficient ARCH model (MvARCH) is proposed and its estimation and corresponding asymptotic property are discussed. Both simulation and real data examples are used for illustration.

3.2 Parametric Conditional Heteroscedasticity Models

3.2.1 Autoregressive Conditional Heteroscedasticity (ARCH) Model

ARCH model was introduced by Engle ([21]) to offer key insight into the distinction between the conditional and unconditional second order moments. The ARCH regression model for a dependent variable y_t is formally defined as:

$$y_t|\mathcal{F}_{t-1} \sim \mathcal{N}(x_t'\beta, h_t),\ h_t = h(\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_{t-p}; \alpha),\ \varepsilon_t = y_t - x_t^{'}\beta,$$

where $\mathcal{F}_t$ is the information set at time t, x_t is a vector of exogenous variables or lagged values of the dependent variables, and β and α are parameter vectors. Application of ARCH has primarily focused on the linear ARCH(p) model given by

$$h_t = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \cdots + \alpha_p\varepsilon_{t-p}^2,\ \alpha_0 > 0,\ \alpha_i \geq 0. \tag{3.1}$$

For the process $\{\varepsilon_t\}$ to be weakly stationary, the parameter in (3.1) must satisfy $\Sigma_{i=1}^p \alpha_i < 1$. Consequently,

$$\text{var}(\varepsilon_t) = \sigma^2 = \frac{\alpha_0}{1 - \Sigma_{i=1}^p \alpha_i},\ h_t = (1 - \Sigma_{i=1}^p \alpha_i)\sigma^2 + \alpha_1\varepsilon_{t-1}^2 + \cdots + \alpha_p\varepsilon_{t-p}^2,$$

where h_t can be regarded as a weighted average of the ‘global’ variance σ^2 and the ‘local’ variances $\varepsilon_{t-1}^2, \cdots, \varepsilon_{t-p}^2$ ([21]). In many empirical applications, it has been shown that ARCH process provides a good fit for a wide variety of financial return time series; see e.g. [5, 9] among others.

3.2.2 Extensions of ARCH Model

Since Engle's paper, many extensions and generalizations have emerged to refine the modeling volatility with specific characteristics; see [22] for a good survey. As pointed out by [39], this research falls into three general categories, each of which addresses one of the assumptions of the linear ARCH specification. The first area of study concerns about the conditional normality assumption, with Student's t ([22]) or the generalized error distribution ([43]) among others proposed to account for heavy tails of the conditional distribution. The second extension of the ARCH are models with nonlinear functional forms for h_t. Some examples are

$$h_t = \exp(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2) \text{ ([22])},$$

$$\log(h_t) = \alpha_0 + \alpha_1 \log(\varepsilon_{t-1}^2) + \cdots + \alpha_p \log(\varepsilon_{t-p}^2) \text{ ([62])}.$$

Note that the parameters need no longer to be nonnegative for h_t to be positive.

The last area, by far the one having received the most amount of attention is that h_t is itself an $ARMA$ process with ϵ_t^2 acting as innovations, which avoid the problem that often a large number of parameters are called in ARCH for better performance. These are the so called Generalized ARCH models([8]), abbreviated as GARCH model. The volatility h_t in the GARCH(p,q) model is given by

$$h_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}, \quad \alpha_0 > 0, \alpha_j \geq 0, \ \beta_j \geq 0, \tag{3.2}$$

where $\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j < 1$ for $\{\varepsilon_t\}$ to be weakly stationary. We can see that GARCH models allow for both a long memory and a much more flexible lag structure. In practice, the most frequently used is the GARCH(1,1) model.

To account for the different responses of the price of a financial asset to positive and negative shocks, i.e. the asymmetric impacts of the 'good news' and the 'bad news' on financial returns, variations of GARCH were introduced with leverage terms included in the expression of h_t. Well-known examples are

1. Exponential GARCH(q,p) model ([60])

$$\ln h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i(\gamma[|\varepsilon_{t-i}| - E|\varepsilon_{t-i}|] + \phi\varepsilon_{t-i}) + \sum_{i=1}^{q} \beta_i \ln h_{t-i};$$

2. Threshold GARCH model ([84, 85])

$$h_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^{p} \gamma_j S_{t-j} \varepsilon_{t-j}^2 + \sum_{i=1}^{q} \beta_i h_{t-i},\ S_{t-j} = I_{\{\varepsilon_{t-j}<0\}}.$$

3.3 Nonparametric Volatility Model

By nonparametric volatility model, we mean that $h_t = \sigma^2(x_t)$, where x_t can be a vector of exogenous variables or lagged values of y_t and $\sigma(.)$ is a nonnegative and smooth function. Contrary to ARCH and GARCH models, nonparametric volatility model puts no restriction on the exact form of h_t, thus allowing more space of flexibility. To accommodate time series with nonzero mean, a more generalized nonparametric mean and volatility model for strictly stationary process $\{(y_t, x_t)\}$ is given by

$$y_t = m(x_t) + \sigma(x_t)\varepsilon_t,\ E(\varepsilon_t|x_t) = 0,\ Var(\varepsilon_t|x_t) = 1, \tag{3.3}$$

while the conditional distribution of ε_t given $x_t = x$ may still depend on x. Let $y_t = (S_{(t+1)\triangle} - S_{t\triangle})/\triangle$, $x_t = S_{t\triangle}$. Then (3.3) is related to the continuous model([29])

$$dS_t = \mu(S_t) + \sigma(S_t)dW_t, \tag{3.4}$$

which was used to model return structure dynamics by [79, 93] .

An obvious and direct estimator for the volatility function in (3.3) is $\hat{\sigma}_d^2(x) = \hat{v}(x) - (\hat{m}(x))^2$, where $\hat{m}(.)$and $\hat{v}(.)$ are estimators for $m(.)$ and $v(x) = E(y_t^2|x_t = x)$ respectively; see [90, 37]. However, as $\hat{\sigma}_d^2(.)$ is not always non-negative and is greatly biased ([29]), [29] suggested the following residual-based estimator of $\sigma(.)$

Step 1. Estimate $m(x_i),\ i = 1, \ldots, n$, by $\hat{a}$ given by

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{j=1}^{n} \{y_j - a - b(x_j - x_i)\}^2 K\{(x_j - x_i)/h_1\}.$$

Step 2. Compute squared residuals $\hat{r}_i = \{y_i - \hat{m}(x_i)\}^2,\ i = 1, \ldots, n.$

Step 3. Obtain the local linear estimator $\hat{\sigma}_r^2(x_i) = \hat{a},\ i = 1, \ldots, n$ by solving

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{j=1}^{n} \{\hat{r}_j - a - b(x_j - x_i)\}^2 K\{(x_j - x_i)/h_2\}, \tag{3.5}$$

where $K(.)$ is a symmetric density function and $h_1,\ h_2$ are two smoothing parameters. This method, as pointed out by [29], performs almost as well as the local linear estimator $\hat{\sigma}_b^2(x)$ when the regression function $m(.)$ is known.

3.4 Parametric or Nonparametric? An Empirical Study

This example concerns the yields of the three-month Treasury Bill from the second market rates on Fridays. The second market rates are annualized using a 360-day year of bank interest and are quoted on a discounted basis. The data, which consists of 1735 weekly observations, from Jan 5, 1962 to Mar 31, 1995, is presented in Fig. 3.1(a). This data has been analyzed by various authors and the complete data set is available at $www.federalreserve.gov/releases/h15/data.htm.$

Let y_t denote the interest rate series. Following the steps in [3, 29], we first fit an AR model using Yule-Walker method with the order selected by AIC criterion

$$\begin{aligned} y_t &= z_t + 1.2641 y_{t-1} - .2766 y_{t-2} + .0444 y_{t-3} + .036 y_{t-4} \\ &\quad -.0459 y_{t-5} - .028 y_{t-6} - .0921 y_{t-7} + .0974 y_{t-8}, \end{aligned}$$

where z_t is the 'residual' of the AR fit. This is different from the $AR(5)$ model in [29, 3], which is caused by the difference in estimation method used for AR fitting. However, we will see later that this does not have significant impact on the estimation of volatility. The 'residual' z_t is plotted against y_{t-1} in Fig. 3.1(b), where the solid line is the Nadaraya-Waston estimator of the mean function $m(x) := E(z_t|y_{t-1} = x)$ and a weakly upward tendency can be noticed up to $y_{t-1} = 14$, which was also observed by [29]. The detrended z_t, i.e. $z_t - \hat{m}(y_{t-1})$ and the residual-based estimator of the conditional variance $\sigma^2(x) := Var(z_t|y_{t-1} = x)$ are illustrated in Fig. 3.1(c) and (d) respectively, which are almost identical to that in [29]. The bandwidth selected by cross-validation is 1.3537 for $\hat{m}(x)$ and 2.6458 for $\hat{\sigma}_r^2(x)$. The overall fitted model is

$$\begin{aligned} y_t &= 1.2641 y_{t-1} - .2766 y_{t-2} + .0444 y_{t-3} + .036 y_{t-4} - .0459 y_{t-5} \\ &\quad -.028 y_{t-6} - .0921 y_{t-7} + .0974 y_{t-8} + \hat{m}(y_{t-1}) + \hat{\sigma}_r(y_{t-1})\varepsilon_t, \qquad (3.6) \end{aligned}$$

with $E(\varepsilon_t|y_{t-1}) = 0,\ Var(\varepsilon_t|y_{t-1}) = 1.$

Among the parametric models, I tried AR(q)+GARCH$(1,1)$ model, i.e.

$$y_t = c_0 + c_1 y_{t-1} + \cdots, + c_q y_{t-q} + \epsilon_t, \quad \epsilon_t \sim GARCH(1,1), \qquad (3.7)$$

with order q chosen by AIC. To compare model (3.6) and (3.7), we compute the exceedance ratio (ER) defined in [27] for performance evaluation. This measure counts

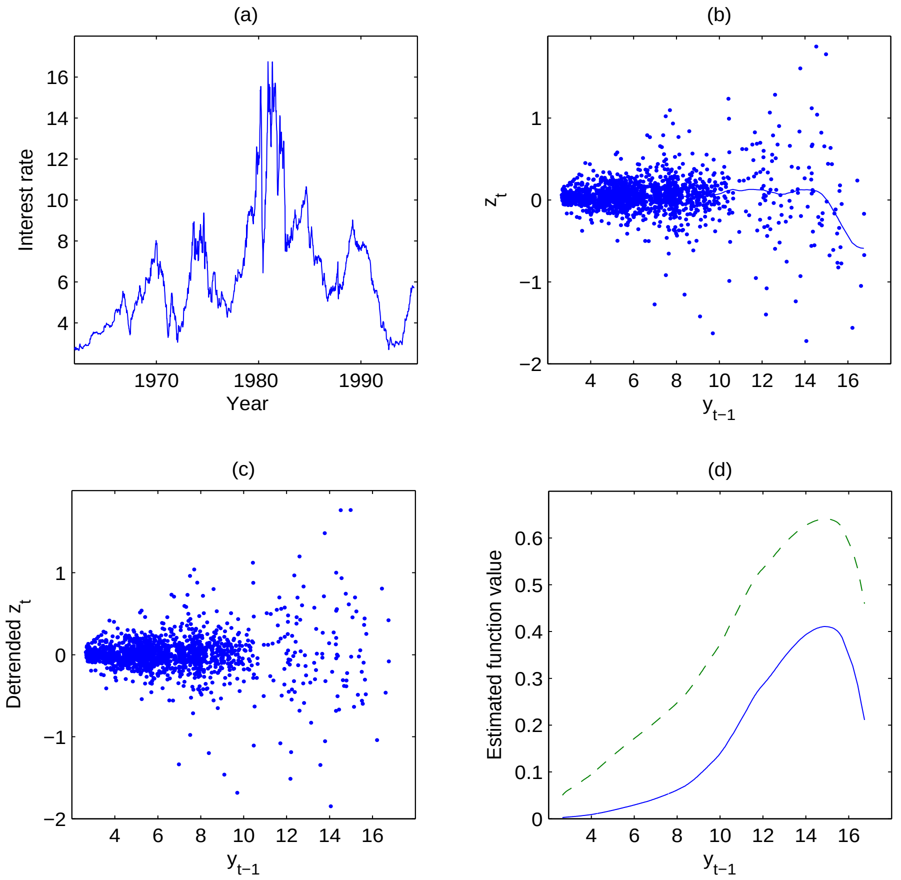

Figure 3.1: *(a) raw data y_t; (b) scatterplot of (y_{t-1}, z_t) and $\hat{m}(y_{t-1})$ (solid line); (c)scatterplot of $(y_{t-1}, z_t - \hat{m}(y_{t-1}))$; (d) $\hat{\sigma}_r^2(y_{t-1})$ and $\hat{\sigma}_r(y_{t-1})$.*

number of events for which the loss of the asset exceeds the loss predicted by the normal model at a given confidence level α. With one-day forward forecasted volatility $\hat{h}_t$ from previous 435 records, the ER with a post sample of size N at time point T is computed as

$$\mathrm{ER}_T = N^{-1} \sum_{t=T-1}^{T-N} I(\epsilon_t < \Phi^{-1}(\alpha)\hat{h}_t^{1/2}). \tag{3.8}$$

Note that ER is closely related to VaR mentioned in Section 1. Fig. 3.2 presents $\mathrm{ER}_T(T = 736, \cdots, 1735)$ with $\alpha = 5\%$. Although neither is satisfactory, ER from (3.6) fluctuate around 5%, while those from (3.7) is always below 2.5%, i.e. the volatility is always over forecasted if we use model (3.7).

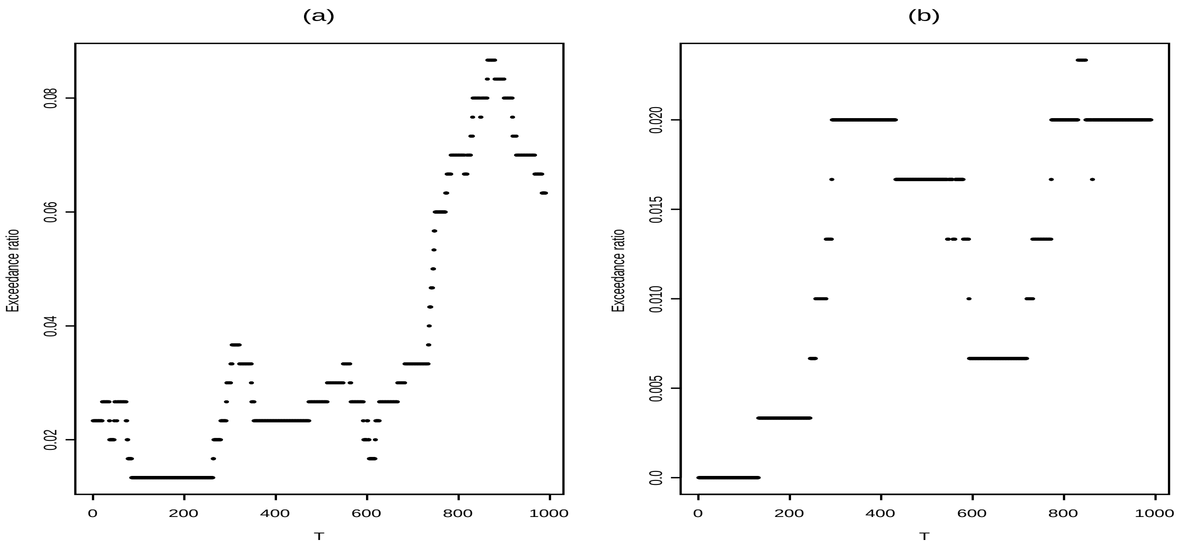

Figure 3.2: *Exceedance ratio with post sample size* 300*: (a) model (3.6); (b) model (3.7).*

3.5 Monotone Constrained Varying Coefficient ARCH Models

3.5.1 Introduction

The parametric volatility models, including GARCH and its extensions are time-homogeneous models, i.e. the parametric structure of the interest process is assumed to be constant throughout the whole sample span. This is a possibly unrealistic assumption, in particular, as far as forecasting is concerned.

Our motivation to introduce varying coefficients comes from the empirical study of [46] and [87]. The data set in [46] comprises daily return and trading volume for 20 actively traded stocks for which options trade on CBOE (Chicago Board Options Exchange). They found that while the GARCH effect is quite strong, as measured by the summation $(\alpha_1 + \beta_1)$ of the fitted GARCH(1,1) model, it tends to disappear if daily trading volume v_t is included in the conditional variance equation, i.e.

$$h_t = \alpha_0 + \beta_1 h_{t-1} + \alpha_1 Y_{t-1}^2 + \alpha_2 v_t. \tag{3.9}$$

Furthermore, they argued that daily trading volume has significant explanatory power regarding h_t, since α_2 in (3.9) is nonzero (in fact positive) at 95% significant level for all 20 stocks. Their findings are not isolated, as previous empirical studies of both futures and equity markets always find a positive association between the return variability h_t and the trading volume; see [14, 23] for possible explanations for such phenomena.

This idea was further developed by [87]. He found that trading volume not only contributes positively to the contemporaneous volatility, but also has a negative impact on the subsequent volatility. Specifically, he considered the following model

$$|Y_t| = \alpha_0 + \alpha_1|Y_{t-1}| + \alpha_2|Y_{t-2}| + \beta_1 TO_t + \beta_2 TO_{t-1} + \epsilon_t, \tag{3.10}$$

where $|Y_t|$ is the absolute value of individual stock daily return and TO_t is the turnover acquired by dividing the traded shares by corresponding shares outstanding. He reported that while β_1 in (3.10) is positive and significant, β_2 is negative and significant. Although (3.10) is specified about the absolute return, it is expected that similar result holds for h_t. This inspired us to conjecture that what really has an impact on volatility h_t is not the contemporary trading volume v_t, but $V_t := v_t - v_{t-1}$, the difference in trading volume between today and the previous day. This combined with the findings of [46] results in the following model

$$h_t = \alpha_0 V_t, \tag{3.11}$$

which, when considered as a linear ARCH model with $p = 0$, implied that the constant term in h_t specified by (3.1) is actually a linear function, or more generally, a monotone function of V_t. Similar generalization of other coefficients $\alpha_j,\ j = 1, \cdots, p$ in (3.1) leads to the Monotone constrained varying coefficient ARCH(p) (MvARCH) model

$$Y_t = \Big\{a_0(V_t) + \sum_{j=1}^{p} a_j(V_t)Y_{t-j}^2\Big\}^{1/2}\epsilon_t,\ E\{\epsilon_t|V_t, \mathcal{F}_1^{t-1}\} = 0,\ Var(\epsilon_t|V_t, \mathcal{F}_1^{t-1}) = 1, \tag{3.12}$$

where $a_i(.),\ i = 0, \cdots, p$ are unknown smooth functions of V_t satisfying

$$a_j(.) \geq 0,\ a_j(v_1) \geq a_j(v_2),\ \text{if } v_1 \geq v_2,\ j = 0, \cdots, p, \tag{3.13}$$

and $\mathcal{F}_1^{t-1}$ is the $\sigma-$algebra generated by $\{V_j, Y_j\}_{j=1}^{t-1}$.

Next, we discuss the geometric ergodicity of $\{Y_t\}$ in (3.12). Note that (3.12) can be transformed into an ordinary varying coefficient regression model by taking squares of both sides. Suppose that every $a_j(.),\ j = 1, \cdots, p$, can be written as $a_j(.) = \alpha_j(.) + \beta_j(.)$ with $\beta_j(v)|v|$ bounded on $\mathcal{R}$ and $|\alpha_j(.)| < c_i$, such that all the roots of $\lambda^p - c_1\lambda^{p-1} - \cdots - c_p = 0$ are inside the unit circle. If the density function of ϵ_t is positive almost everywhere and $\lim_{|v|\to\infty} \sup |a_0(v)/v| \to 0$, then $\{Y_t\}$ is geometrically ergodic; see [12]. By the results in [63], a geometrically ergodic time series is a strongly mixing sequence. Therefore, it is safe to impose the strongly mixing assumption on model (3.12) under the aforementioned conditions.

Constraint (3.13) complicates the estimation of (3.12), since the function value at any given points are no longer determined locally but associated with one another. Various techniques for estimating constrained nonparametric functions have been developed; see e.g. [51, 67]. In theory, it is possible to incorporate the constraint into all kinds of smoothing methods, but arguably it is not as convenient to do so with kernel-type smoothing as with spline-based smoothing methods; see the comments of Wahba on [66]. Estimation of (3.12)-(3.13) is based on local linear smoothing and coincides in some sense with the 'globalization' approach of [53]; see also [51]. The proposed method will enjoy the same convenience as spline-based approach in terms of incorporating constraints and might be appealing for users who prefer kernel-type smoothing.

3.5.2 Globalization Kernel Smoothing Estimation

Let $\xi_t = \big\{a_0(V_t) + \sum_{j=1}^{p} a_j(V_t)Y_{t-j}^2\big\}(\epsilon_t^2 - 1)$. Then the weakly stationary process given by (3.12) can be written as a bona fide varying coefficient model

$$Y_t^2 = a_0(V_t) + \sum_{j=1}^{p} a_j(V_t)Y_{t-j}^2 + \xi_t. \tag{3.14}$$

To start with, we adopt the local linear smoothing approach to estimate the coefficient functions in (3.14). Local linear approximation method is chosen here because of its high statistical efficiency in an asymptotic minimax sense and design-adaptive property ([24]), besides the capability of automatically correcting the edge effects ([25, 38, 70]). The basic idea is to treat the value of the function $a_k(.)$ at any given point, v_0 say, as a local parameter and to approximate $a_k(.)$ by a local linear function

$$a_k(V) \simeq a_k(v) + a_k'(v)(V - v), \ \ k = 0, 1, \cdots, p, \tag{3.15}$$

for V in a neighborhood of v. The local linear estimator of $\mathbf{a}(v) = (a_0(v), \cdots, a_p(v))^\top$ and $\mathbf{a}'(v) = (a_0'(v), \cdots, a_p'(v))^\top$ are given by $\hat{\mathbf{a}}(v) := (\hat{a}_0, \cdots, \hat{a}_p)^\top$ and $\hat{\mathbf{a}}'(v) := (\hat{b}_0, \cdots, \hat{b}_p)^\top$ respectively, where $(\hat{a}_0, \hat{b}_0, \cdots, \hat{a}_p, \hat{b}_p)^\top$ is the solution to the minimization problem

$$\min_{\substack{a_0,\cdots,a_p,\\ b_0,\cdots,b_p}} \sum_{i=1}^{n} \Big\{ Y_i^2 - \sum_{k=0}^{p} (a_k + b_k V_{iv}) Y_{ik}^2 \Big\}^2 K_h(V_{iv}),$$

where $V_{iv} := V_i - v,\ Y_{i0} := 1,\ Y_{ik} := Y_{i-k},\ K$ is an univariate density function, h is the smoothing parameter and $K_h(.) := K(./h)/h$. For details about local linear smoothing, see [26]. If the marginal density of V, $f_V(v)$ is positive, then under some regularity conditions (see Appendix C), [11] proved that

$$(nh)^{1/2} \Big\{ \hat{\mathbf{a}}(v) - \mathbf{a}(v) - \frac{\mu_2}{2} h^2 \mathbf{a}_j''(v) \Big\} \to N(0, \Theta_1(V)), \tag{3.16}$$

where $\Theta_1(v) := \mu_0^* f_V^{-1}(v) \Omega^{-1}(v) \Omega^*(v) \Omega^{-1}(v),\ \mathbf{a}_j''(v) = (a_0''(v), \cdots, a_p''(v))^\top$ and the definition of $\Omega(v),\ \Omega^*(v)$ and μ_0^* are given in Appendix C.

As the estimated function is often not monotonic, the globalization kernel smoothing method is needed to solve this problem by estimating the values of $a_k(.)$ at desired points simultaneously. Let $v^1 < v^2 < \cdots < v^m$ denote m equally spaced points on $\mathcal{D}$, the compact support of V_t. Let $X_j = (Y_{j0}, Y_{j1}^2, \cdots, Y_{jp}^2)^\top,\ j = 1, \cdots, n$, where $Y_{j0} \equiv 1$ and $Y_{jk} = Y_{j-k},\ k = 1, \cdots, p$. For $i = 1, \cdots, m$, set

$$\tilde{X}_i = \begin{pmatrix} X_1^\top \otimes (1, V_{1i}) \\ \cdots \\ X_n^\top \otimes (1, V_{ni}) \end{pmatrix},\ \mathcal{X} = \mathrm{diag}(\tilde{X}_1, \cdots, \tilde{X}_m),\ \mathcal{Y} = \mathbf{1}_{m\times 1} \otimes \begin{pmatrix} Y_1^2 \\ \cdots \\ Y_n^2 \end{pmatrix}, \tag{3.17}$$

$$\mathbf{1}_{m\times 1} = (1, \cdots, 1)^\top,\ W_i = \mathrm{diag}\{K_h(V_{1i}), \cdots, K_h(V_{ni})\},\ W = \mathrm{diag}(W_1, \cdots, W_m),$$

$$\beta = (a_{1,0}, b_{1,0}, a_{1,1}, b_{1,1}, \cdots, a_{1,p}, b_{1,p}, \cdots, a_{m,0}, b_{m,0}, a_{m,1}, b_{m,1} \cdots, a_{m,p}, b_{m,p})^\top,$$

where $V_{ji} := V_j - v^i$ and $\otimes$ denotes the Kronecker product. Let $e_{k,m}$ be the $m \times 1$ vector with 1 at the kth position and others 0. Denote by $\hat{\alpha}$ the solution to

$$\min_{\beta} (\mathcal{Y} - \mathcal{X}\beta)^{\top} W (\mathcal{Y} - \mathcal{X}\beta). \tag{3.18}$$

Then $\mathbf{a}(v^i)$ and $\mathbf{a}'(v^i),\ i = 1, \cdots, m$ can be estimated by $\hat{\mathbf{a}}(v^i) := (\hat{a}_{i,0}, \cdots, \hat{a}_{i,p})^{\top}$ a nd $\hat{\mathbf{a}}'(v^i) := (\hat{b}_{i,0}, \cdots, \hat{b}_{i,p})^{\top}$ respectively, where

$$\hat{a}_{i,k} = e^{\top}_{(2p+2)(i-1)+2k+1,2m(p+1)} \hat{\alpha}, \quad \hat{b}_{i,k} = e^{\top}_{(2p+2)(i-1)+2k+2,2m(p+1)} \hat{\alpha}.$$

To reflect the constraint (3.13), what we need is the solution to (3.18) subject to

$$a_{j,k} \le a_{j+1,k},\ j = 1, \cdots, m-1,\ 0 \le k \le p. \tag{3.19}$$

To do this, first rewrite (3.18) as

$$\min_{\beta} (B + \beta^{\top} Q \beta - 2C^{\top}\beta), \tag{3.20}$$

where $Q = \mathcal{X}^{\top} W \mathcal{X},\ B = \mathcal{Y}^{\top} W \mathcal{Y},\ C = \mathcal{X}^{\top} W \mathcal{Y}$. Let A be a $(m-1)(p+1)$ by $2m(p+1)$ matrix, with $A(j, 2j-1) = -1,\ A(j, 2j+2p+1) = 1,\ j = 1, \cdots, (m-1)(p+1)$, and other entries zero. Then (3.20) subject to (3.19) is equivalent to

$$\min_{\beta} (B + \beta^{\top} Q \beta - 2C^{\top}\beta) \quad \text{subject to } A\beta \ge \mathbf{0}, \tag{3.21}$$

which is a quadratic minimization programming subject to inequality constraints.

Let $\tilde{\alpha} = (\tilde{a}_{1,0}, \tilde{b}_{1,0}, \tilde{a}_{1,1}, \tilde{b}_{1,1}, \cdots, \tilde{a}_{1,p}, \tilde{b}_{1,p}, \cdots, \tilde{a}_{m,0}, \tilde{b}_{m,0}, \tilde{a}_{m,1}, \tilde{b}_{m,1} \cdots, \tilde{a}_{m,p}, \tilde{b}_{m,p})^{\top}$ be the solution to (3.21) and $\tilde{\mathbf{a}}(v^i) := (\tilde{a}_{i,0}, \cdots, \tilde{a}_{i,p})^{\top},\ i = 1, \cdots, m$, and $\delta_n = (nh/\ln n)^{-1/2}$.

Theorem 3.1 *Suppose* $(A1) - (A8)$ *in Appendix C hold. Then*

$$\sup_{v \in \mathcal{D}} |\hat{a}_j(v) - a_j(v)| = O_p(h^2 + \delta_n).$$

For the strong uniform convergence rate and asymptotic normality of $\tilde{a}_j(v^i)$, we have

Theorem 3.2 *If (3.13) and* $(A1) - (A9)$ *in Appendix C hold, then*

$$\sup_{1\le i\le m} \left|\tilde{a}_j(v^i) - a_j(v^i)\right| = O_p(h^2 + \delta_n). \tag{3.22}$$

Let $\alpha := (\mathbf{a}(v^1)^\top, \mathbf{a}'(v^1)^\top, \cdots, \mathbf{a}(v^m)^\top, \mathbf{a}'(v^m)^\top)^\top$. *If* $A\alpha > \mathbf{0}$, *then*

$$(nh)^{1/2}\Big\{\tilde{\mathbf{a}}(v^i) - \mathbf{a}(v^i) - \frac{\mu_2}{2}h^2\mathbf{a}_j^{''}(v^i)\Big\} \to N(0, \Theta_1(v^i)) \ i = 1, \cdots, m. \tag{3.23}$$

If there exist matrices $A_1,\ A_2$ *such that* $A^\top = [A_1^\top|A_2^\top]$ *and* $A_1\alpha > \mathbf{0},\ A_2\alpha = \mathbf{0}$, *then*

$$(nh)^{1/2}\Big\{\tilde{\mathbf{a}}(v^i) - \mathbf{a}(v^i) - \theta(v^i)h^2\mathbf{a}_j^{''}(v^i)\Big\} \to N(0, \Theta_2(v^i)),\ i = 1, \cdots, m, \tag{3.24}$$

for some vector $\theta(v^i)$ *and matrix* $\Theta_2(v^i)$ *given in the proof.*

Similar consistency issues were addressed in [49] for linear regression models, where a closed form of the estimator was given in the situation that the constraints can be identified as strict inequality or equality. Theorem 3.2 considers not only the strong uniform convergence, but also the asymptotic normality for the constrained estimator. We can see that if the inequality in (3.19) holds strictly, estimators with and without constraints share common limit distribution, as $n \to \infty$.

3.6 Simulation Results

In this section, the performance of estimators from (3.20) is compared with from (3.21). Because of the presence of heteroscedasticity, it is appropriate to use weights other than W defined in (3.18), which treats ξ_t's in (3.14) as if they were conditionally homoscedastic. To examine the influence of the weight on the performance of the estimator, besides

W (labeled 'U'), two other ways to decide the weight function are also considered. The first is $W * diag(h_1^{-2}, \cdots, h_n^{-2}, \cdots, h_1^{-2}, \cdots, h_n^{-2})$, i.e. as if the true volatility is known. Generally speaking, this should be the optimal weight (labeled 'T'). The second way is through iteration (labeled 'I'). That is, starting with W, each time we get the estimated $\hat{h}_t,\ t = 1, \cdots, n$, the weight is updated as $W * diag(\hat{h}_1^{-2}, \cdots, \hat{h}_n^{-2}, \cdots, \hat{h}_1^{-2}, \cdots, \hat{h}_n^{-2})$. We find empirically that the results usually become stable after five or six iterations.

Example 3.3 Consider the nonlinear time series model

$$Y_i = \left\{a_0(V_i) + a_1(V_i)Y_{i-1}^2\right\}^{1/2}\epsilon_i,$$

where $\{V_i\}$ and $\{\epsilon_i\}$ are two independent sequences of independent random variables with $V_i \sim U[0,3]$, $\epsilon_i \sim N(0,1)$ and

$$a_0(V) = \frac{\exp\{(V+2)^3/30\}}{3\{9 + 2\exp(V)\}}, \quad a_1(V) = \frac{2}{3}\{1 - \exp(-2V)\}.$$

For each simulated sample, the performance of estimators both with and without constraints is evaluated based on the mean absolute deviation error(MAD),

$$MAD_k = \frac{1}{n_{grid}} \sum_{j=1}^{n_{grid}} |\hat{a}_k(v^j) - a_k(v^j)|,\ k = 0, 1,$$

where $\{v^j,\ j = 1, \cdots, n_{grid}\}$ are grid points on $[0,3]$ with $n_{grid} = 99$. Results from 200 simulations with $n = 800$ are summarized as Fig. 3.3, where the upper two panels are the box plots of MAD_0 and MAD_1 respectively. The columns are labeled in the way such that the first letter 'C'/'W' indicates with/without constraints and the second letter 'U'/'T'/'I' indicates the type of weight used. First we can see that the performance of estimator after iteration is comparable to that when the true volatility is known. As anticipated, estimators from (3.21) performs uniformly better than from (3.20). The lower two panels are the empirical pointwise 90% percentiles for $a_0(V)$ and $a_1(V)$ respectively, which show that estimator with constraints lies in a slightly narrower neighborhood around the true value than that without constraint does. Results regarding a typical simulated data set are presented in Fig. 3.4, which leads to similar conclusion as Fig. 3.3.

To examine the effect of correlation on estimation performance, we simulate $\{V_i\}$ from an AR(2) model $V_i = -0.4V_{i-1} + 0.3V_{i-2} + \varepsilon_i$, where $\{\varepsilon_i\}$, independent of $\{\epsilon_i\}$, is a sequence of independent $N(0, 0.01)$ random variables and $a_1(V) = \frac{2}{3}\{1 - \exp(-2V - 0.8)\}$. MAD_0 and MAD_1 are calculated on grid points of $[-0.4, 0.3]$ with $n_{grid} = 69$. Box plots based on 200 simulations are given in Fig. 3.5.

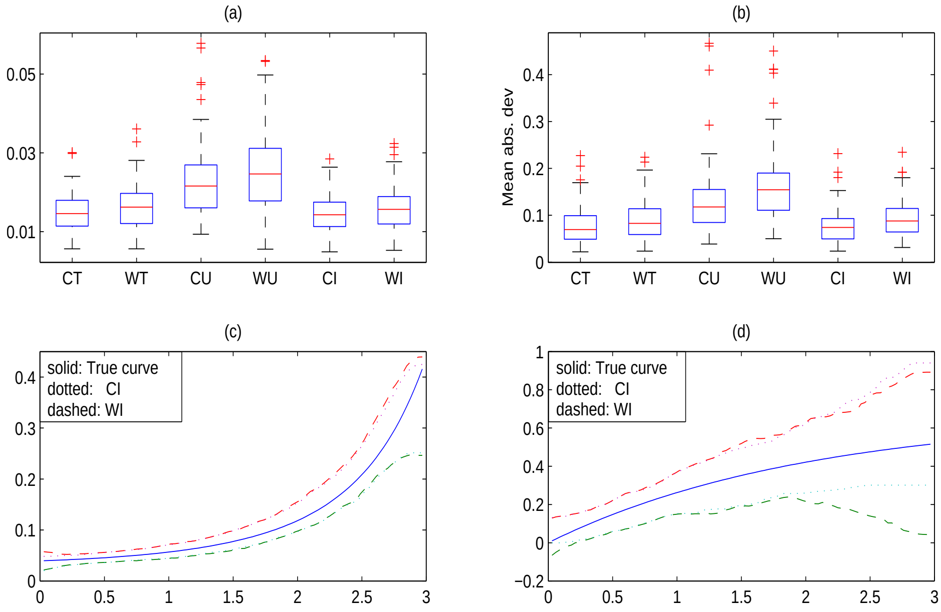

Figure 3.3: *Box plots of MAD_0 and MAD_1 and pointwise 90% percentile.*

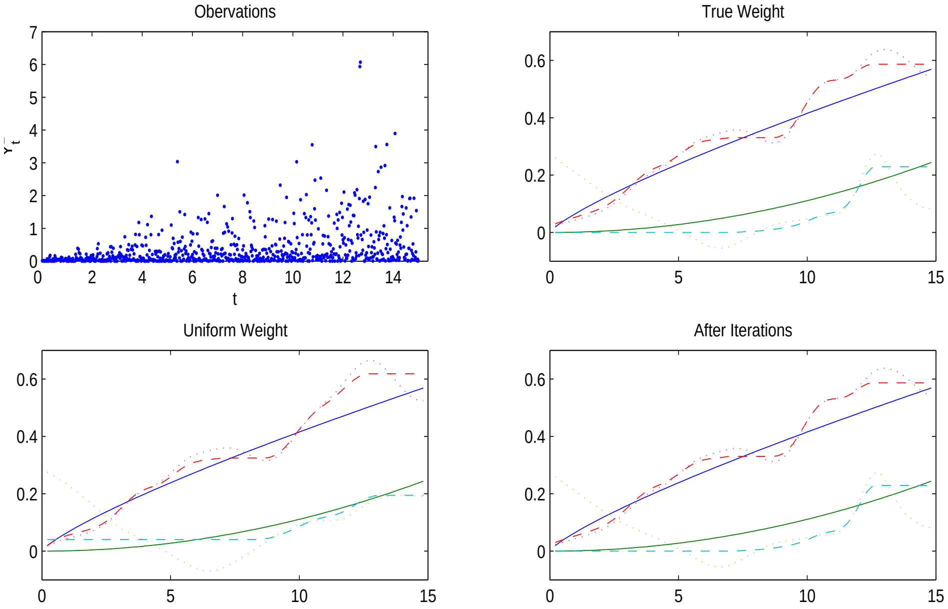

Figure 3.4: *Solid: true curves; dashed, CI; dotted, WI.*

3.7 Empirical Study

The original data consist of daily observations of closing price P_t and trading volume v_t of three stocks with tickers IBM, BP, and GM. A brief description of the three data sets is given in Table 3.7. In Fig. 3.6, panels on the left are plots of series of daily return $r_t := \log(P_t/P_{t-1})$ and those on the right are plots of the differenced trading volume $V_t := v_t - v_{t-1}$.

The performance comparison is based on the exceedance ratio (ER) defined in (3.8) with confidence level $\alpha = 0.01$. One-day forward forecasted volatility $\hat{h}_t$ is calculated based on immediately previous 800 records and the post sample size of ER is 300. Besides

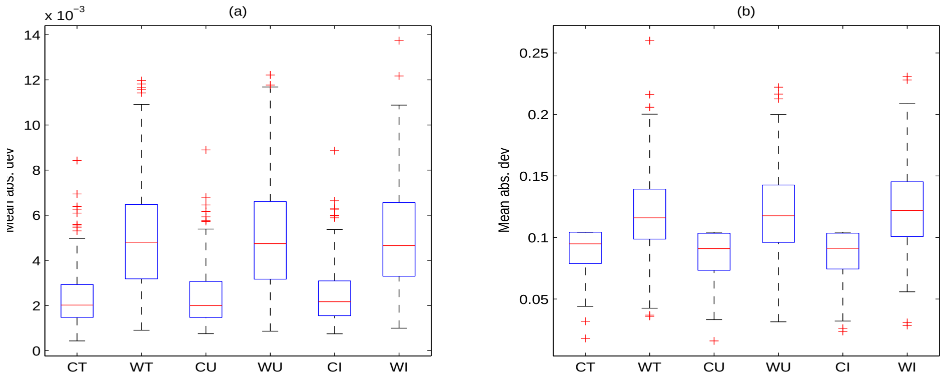

Figure 3.5: *Box plots of MAD_0 and MAD_1 with correlated V_t.*

GARCH(1,1) model, the following models are also considered

$$h_t = a_0(V_t) + \sum_{k=1}^{p} a_k(V_t) r_{t-k}^2, \;\; a_k(.) \geq 0, \;\; k = 0, \cdots, p. \tag{M1}$$

$$h_t = a_0(|V_t|) + \sum_{k=1}^{p} a_k(|V_t|) r_{t-k}^2, \;\; a_k(.) \geq 0, \;\; k = 0, \cdots, p. \tag{M2}$$

$$h_t = a_0(V_t) + \sum_{k=1}^{p} a_k(V_t) r_{t-k}^2, \;\; a_k(.) \geq 0 \text{ and increasing}, \; k = 0, \cdots, p. \tag{M3}$$

$$h_t = a_0(|V_t|) + \sum_{k=1}^{p} a_k(|V_t|) r_{t-k}^2, \;\; a_k(.) \geq 0 \text{ and increasing}, \; k = 0, \cdots, p. \tag{M4}$$

We find that with the same order p, model (M2) outperforms (M1) and model (M4) delivers better results than (M3). The reason that $|V_t|$ is more powerful than V_t in explaining volatility may be that, big changes of either sign in trading volume are always accompanied by big changes in price. Therefore, from now on we focus on the study of GARCH(1,1), model (M2) and (M4). For model (M2) and (M4), the order p which gives the best results in terms of ER for each data set are given in the last column of Table 3.7. Panels on the left hand side of Fig. 3.7 depict ER of GARCH(1,1) and model (M4), while those on the right hand side present ER of GARCH(1,1) and model (M2).

Table 3.1: Details of Three Stocks Data Sets

Ticker	Country	Period	Sample size	Order
IBM	USA	Dec 29, 1975 - July 26, 1996	5202	6
BP	UK	Mar 9, 1978 - Oct 11, 1996	4902	4
GM	USA	July 15, 1979 - Dec 1, 1998	4702	2

Through comparison of the graphs on the left with those on the right, it is clear that the presence of monotonicity constraint significantly enhances the accuracy of volatility prediction. Focusing on panels on the left, we can see that MvARCH model performs as well as, and sometimes better better than GARCH$(1, 1)$ model.

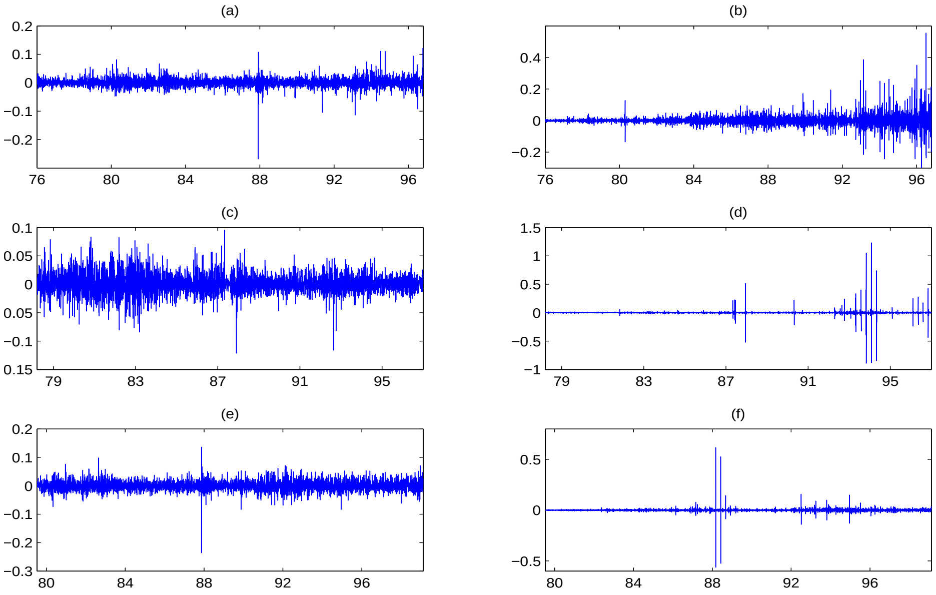

Figure 3.6: *Plots of r_t and V_t: IBM,(a)-(b);BP,(c)-(d);GM,(e)-(f).*

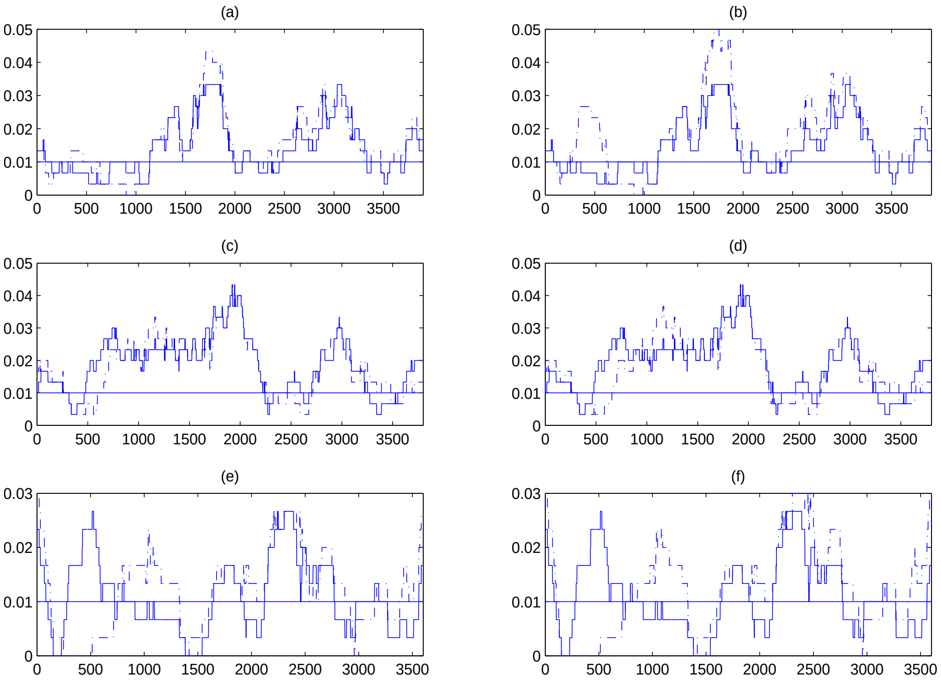

Figure 3.7: *ER: '–', GARCH(1,1); '-.', MvARCH. IBM,(a)-(b);BP,(c)-(d);GM,(e)-(f).*

Bibliography

[1] Akaike, H. (1974) A new look at statistical model identification. *IEEE Trans. Auto. Cont.* **19**, 716-23.

[2] Altham, P.M.E. (1984) Improving the precision of estimation by fitting a model. *J. R. Statist. Soc.* B **46**, 118-9.

[3] Andersen, T.G. & Lund, J. (1997) Estimating continuous time stochastic volatility models of the short term interest rate. *J. Econometric.* **77**, 343-77.

[4] Auestad, B. & Tjϕstheim, D. (1990) Identification of nonlinear time Series: first ordercharacterization and order determination. *Biometrika* **77**, 669-87.

[5] Baillie, Richard T. & Tim Bollerslev (1989) The message in daily exchange rates: a conditional variance tale. *J. Busi. Economet. Statist.* **7**, 60-8.

[6] Barnett, W. A., Powell, J. L. & G. Tauchen (1991) *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* New York: Campbridge University Press.

[7] Bickel, P. J., C. A. J. Klaassen, Y. Ritov, & TJ. A. Wellner (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Jonhns Hopkins Unversity Press.

[8] Bollerslev, Tim (1986) Generalized autoregressive conditional heteroscedasticity. *J. Econometrics* **31**, 307-27.

[9] Bollerslev, Tim (1987) A conditionally heteraskedastic time series model for speculative prices and rates of return. *Rev. Economet. Statist.* **69**, 542-47.

[10] Burman, P. (1989) A comparative study of ordinary cross-validation, v-fold cross-validati on and the repeated learning testing methods. *Biometrika* **76**, 503-14.

[11] Cai, Z., Fan, J. & Yao, Q. (2000) Functional-coefficient regression models for non-linear time series. *J. Am. Statist. Assoc.* **5**, 941-56.

[12] Chen, R. & Tsay, R. S. (1993) Functional-coefficient autoregressive models. *J. Am. Statist. Assoc.* **88**, 298-308.

[13] Cheng, B. & Tong, H. (1993) On residual sums of squares in non-parametric autoregression. *Stochastic Process. Appl.* **48**, 157-74.

[14] Clark, P.K. (1973) A subordinate stochastic process model with finite variance for speculative prices. *Econometrica* **41**, 135-55.

[15] Cottle, R.W., & Dantzig, G.B. (1974) Complementary pivot theories of mathematical programming. In *Studies of Optimization* Vol. **10**, Ed. G.B.Dantzig and B.C.Eaves. Washington: Mathematical Association of America.

[16] Craven, P. & Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics* **31**, 377-403.

[17] D. Bosq (1998) *Nonparametric Statistics for Stochastic Processes,* Lecture Notes in Statistics 110. New York: Springer.

[18] Dantzig, G.B. & Cottle, R.W. (1967) Positive (semi-) definite matrices and mathematical programming. In *Nonlinear Programming*, 55-73, Ed. J. Abadie. Amsterdam: Northholland Publishing Co.

[19] Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Statist. Assoc.* **78**, 316-31.

[20] Efron, B.(1986) How bias is the apparent error rate of a prediction rule. *J. Am. Statist. Assoc.* **81**, 461-70.

[21] Engle, R.F. (1982) Autoregressive conditional heteraskedasticity with estimates of the variance of U.K. Inflation. *Econometrica* **50**, 987-1008.

[22] Engle, R.F. & Tim Bollerslev (1986) Modeling the persistence of conditional variances. *Economic Reviews* **5**, 1-50.

[23] Epps, T. W. & M. L. Epps (1976) The stochastic dependence of security price changes and transaction volumes: implications for the mixture-of-distribution hypothesis. *Econometrica* **44**, 305-21.

[24] Fan, J. (1993) Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21**, 196-216.

[25] Fan, J. & Gijbels, I. (1992) Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008-36.

[26] Fan, J. & Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

[27] Fan, J. & Gu, J. (2003) Semiparametric estimation of value-at-risk. *J. Econometrics* **6**, 261-90.

[28] Fan, J. & Li, R. (1999) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348-60.

[29] Fan, J. & Yao, Q. (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645-60.

[30] Fan, J. & Zhang, W.Y. (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scan. J. Statist.* **27**, 715-31.

[31] Gunst, G.F. & Mason, R.L. (1980) *Regression Analysis and Its Application: A Data-Oriented Approach.* New York: Marcel Dekker, Inc.

[32] Hall, P. kay, J. W. & Titterington D. M. (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521-8.

[33] Hannan, E. J. & Quinn, B. G. (1979) The determination of the order of an autoregession. *J. R. Statist. Soc.* B **41**, 190-5.

[34] Härdle, W., Hall, P. & Ichimura, H. (1993) Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-78.

[35] Härdle, W. & Simar, L. (2003) *Applied Multivariate Statistical Analysis.* Berlin: Springer.

[36] Härdle, W. & Stoker, T. M. (1989) Investigating smooth multiple regression by method of average derivatives. *J. Am. Statist. Assoc.* **84**, 986-995.

[37] Härdle, W. & Tsybakov, A. (1997) Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics.* **81**, 223-42.

[38] Hastie,T. & Loader, C. (1993) Local regression: automatic kernel carpentry (with Discussion). *Statist. Sci.* **8**, 120-43.

[39] Higgins, M.L. & A.K. Bera (1992) A class of nonlinear ARCH models. *International Econometric Review* **33**, 137-58.

[40] Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293-325.

[41] Horowitz, J. L. & Härdle, W. (1996) Direct semiparametric estimation of single-index models with discrete covariates. *J. Am. Statist. Assoc.* **91**, 1632-40.

[42] Hristache, M., Juditski, A. & Spokoiny, V. (2001) Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29**, 595-623.

[43] Hsieh, DA (1989) Modeling heteroscedasticity in daily foreign-exchange rates. *J. Busi. Economet. Statist.* **7**, 307-17.

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

[45] Klein, R.W. & Spady, R.H. (1993) An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387-421.

[46] Lamoureux, C.G. & Lastrapes, W.D. (1990) Heteroskedasticity in stock return data: volume versus GARCH effects. *J. Finance* **1**, 220-9.

[47] Lemke, C.E. (1962) A method of solution for quadratic programs. *Management Science* **8**, 442-53.

[48] Li, K.C. (1991) Sliced inverse regresson for dimension reduction(with Discussion). *J. Am. Statist. Assoc.* **86**, 316-42.

[49] Liew, C.K. (1976) Inequality constrained least-squares estimation. *J. Am. Statist. Assoc.* **71**, 746-51.

[50] Mallows, C. (1973) Some Comments on Cp. *Technometrics* **15**, 661-75.

[51] Mammen, E., Marron, J.S., Turlach, B.A. & Wand, M. P. (2001) A general projection framework for constrained smoothing. *Statist. Sci.* **12**, 232-48.

[52] Mandelbrot B. (1963) The Variation of Certain Speculative Prices. *J. Business* **36**, 394-419.

[53] Marron, J. S. & Chung, S. S. (1997) Presentation of smoothers: the family approach. Unpublished manuscript.

[54] Masry E. (1996) Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571-99.

[55] Masry E. & Tjϕstheim, D. (1995) Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory* **11**, 258-89.

[56] Masry E. & Tjϕstheim, D. (1997) Additive nonlinear ARX time series and projection estimates. *Econometric Theory* **13**, 214-52.

[57] McQuarrie, A.D.R. & C-L, Tsai (1998) *Regression and Time Series model Selection.* Singapore: World Scientific.

[58] Miller, A. J. (2002) *Subset selection in regression (2nd edition).* Chapman and Hall / CRC Press, London and New York.

[59] Naik, P. A. & C-L Tsai (2001) Single-index model selection. *Biometrika* **88**, 821-32.

[60] Nelson, D.B. (1991) Condtitional heteroskedasticity in asset retruns: a new approach. *Econometrica* **59**, 347-70.

[61] Nishiyama, Y. & Robinson, P. M. (2005) The bootstrap and the Edgeworth correction for semiparametric averaged derivatives. *Econometrica* **73**, 903-48.

[62] Pantula, S. G. (1986) Comment on "Modeling the persistence of conditional variances," by R. F. Engle and T. Bollerslev. *Econometric Review* **5**, 71-3.

[63] Pham, D. T. (1986) The mixing property of bilinear and generalized random coefficient autoregressive models. *Stochastic Process. Appl.* **23**, 291-300.

[64] Powell, J. L. (1994) Estimation of semiparametric models. *Handbook of Econometrics,* vol 4, Eds. R. F. Engle and D. F. McFadden. Amsterdam: Elsevier.

[65] Powell, J. L., Stock, J. H. & Stoker, T. M. (1989) Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-30.

[66] Ramsay, J. O. (1988) Monotone regression splines in action (with comments) *Statist. Sci.* **3**, 425-61.

[67] Ramsay, J. O. (1998) Estimating smooth monotone functions. *J. R. Statist. Soc.* B **60**, 365-75.

[68] Rao, C. R. & Wu, Y. (1989) A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369-74.

[69] Robinson, P.M. (1988) Root-n-consistent semi parametric regression. *Econometrica* **56**, 931-54.

[70] Ruppert, D. & Wand, M. P. (1994) Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346-70.

[71] Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

[72] Seifert, B. & Gasser, T. (1996) Finite sample variance of local polynomials: analysis and solutions. *J. Am. Statist. Assoc.* **91**, 267-75.

[73] Shao, J. (1993) Linear model selection by cross-validation. *J. Am. Statist. Assoc.* **88**, 486-94.

[74] Shao, J. (1996) Resampling methods in sample surveys (with discussions). *Statistics* **27**, 203-54.

[75] Shibata, R. (1981) An optimal selection of regression variables. *Biometrika* **68**, 45-53.

[76] Shibata, R. (1984) Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43-9.

[77] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

[78] Speckman, P. (1988) Kernel smoothing in partial linear models. *J. R. Statist. Soc.* B **50**, 413-36.

[79] Stanton, R. (1998) A nonparametric model of term structure dynamics and the market price of interest rate risk. *J. Finance* **52**, 1973-2002.

[80] Stone, M. (1974) Cross-validatoty choice and assessment of statistical prediction (with Discussion). *J. R. Statist. Soc.* B **36**, 111-47.

[81] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267-88.

[82] Tjøstheim, D. & Auestad, B. H. (1994) Nonparametric identification of nonlinear time series: selecting significant lags. *J. Am. Statist. Assoc.* **89**, 1410-9.

[83] Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24-36.

[84] Tong, H. (1978) On a threshold model. In *Pattern Recognition and Signal Processing*, Eds. C.H.Chen. Amsterdam: Sijhoff & Noordhoff.

[85] Tong, H. (1990) *Non-Linear Time Series: A Dynamical System Approach.* Oxford: Oxford University Press.

[86] Von Mises, R. (1947) On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18**, 301-48.

[87] Wang, H.F. (2004) Dynamic volume-volatility relation. Unpublished manuscript.

[88] World Health organization, Reports of a WHO/HEI working group. Bonn, Germany.

[89] Yao, Q. & Tong, H. (1994) On subset selection in non-parametric stochastic regression. *Statistica Sinica* **4**, 51-70.

[90] Yao, Q. & Tong, H. (1994) Quantifying the inference of initial values on nonlinear prediction. *J. R. Statist. Soc.* B **56**, 701-25.

[91] Xia, Y. & Tong, H. (2006) On efficiencies of estimations for single index models. In *Frontiers in Statistics,* Ed. J. Fan and K. Hourl. New York: World Scientific Publishing.

[92] Xia, Y., Tong, H., Li, W. K. & Zhu, L. (2002) An adaptive estimation of dimension reduction space (with Discussion). *J. R. Statist. Soc.* B **64**, 363-410.

[93] Y. Aitsahalia (1996) Nonparametric pricing of interest-rate derivative securities. *Econometrica* **64**, 527-60.

[94] Zhang, P. (1993) Model selection via multifold cross validation. *Ann. Statist.* **21**, 299-313.

Appendix A

Conditions and Proofs for Chapter 1

We impose the following regularity conditions.

(A1) $K(.)$ is a symmetric density function with $K(0) = 1$ and $K(t)(1 + |t|^{2+\delta_1}) \leq M$ for some $\delta_1 > 0,\ M > 0$.

(A2) Bandwidth $h \to 0, nh^6 \to 0, nh^2 \to \infty$.

(A3) For any α and k with $\alpha \cup k \supseteq \alpha_0$, $E(\mathbf{x}_\alpha | \mathbf{x}_k = t)$ has bounded second order derivative with finite second moments and $\Phi_{\alpha,k} := E\left[\{\mathbf{x}_\alpha - E(\mathbf{x}_\alpha|\mathbf{x}_k)\}\{\mathbf{x}_\alpha - E(\mathbf{x}_\alpha|\mathbf{x}_k)\}^\top\right]$ is positive definite.

(A4) The density function $f_k(.)$ of $\mathbf{x}_k$ is bounded away from zero with second-order derivative $|f_k''(.)| \geq \delta_2$, for some $\delta_2 > 0$ over its compact support, for all $1 \leq k \leq p$.

(A1)-(A4) are imposed by [69] to prove the asymptotic normality property of $\hat{\beta}_\alpha$ in partially linear models. Note that (A4) can be relaxed by introducing a trim function in the definition of cross validation function to tackle 'small' random denominator; see [69]. In the notation below, the α in subscript of $\mathbf{x}$ or $\mathbf{x}_i$ are dropped for ease of exposition.

$$K_{ij} = K_h(z_i - z_j),\ w_{ij} = K_{ij} / \sum_{l=1}^{n} K_{il},\ \tilde{\epsilon}_i = \sum_{j=1}^{n} w_{ij}\epsilon_j,\ \tilde{g}_i = \sum_{j=1}^{n} w_{ij} g_j,$$
$$g_i = g(z_i),\ f_i = f(z_i),\ \tilde{y}_i = \tilde{y}(z_i),\ \tilde{x}_i = \tilde{x}(z_i),\ \ e_i = (z_i, \epsilon_i).$$

It follows directly from (A4) that

$$\max_{i,j} w_{ij} = O_p(n^{-1}h^{-1}),\ (1 - w_{ii})^{-2} = 1 + 2w_{ii} + O(w_{ii}^2). \tag{A.1}$$

A statistic V_n is called a $V-$statistic of dimension $k (\geq 1)$ based on i.i.d. observations $\{X_i\}_{i=1}^n$, if it admits the following form

$$V_n = \frac{1}{n^k} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} H(X_{i1}, \cdots, X_{ik})$$

where $H(.)$ is the kernel function and is symmetric in its k arguments. Let $\theta := EH(X_1, \cdots, X_k)$, $H_1(X_1) := E\{H(X_1, \cdots, X_k)|X_1\}$, and $\sigma_1^2 := VarH_1(X_1)$. The following two lemmas are proved in [86] and [69] respectively.

Lemma A.1 *If* $\sigma_1^2 > 0$, *then* $n^{1/2}(V_n - \theta) \to N(0, k^2\sigma_1^2)$, *as* $n \to \infty$.

Lemma A.2 *Under* $(A1) - (A4)$, *we have*

$$(1) \sum_{i=1}^{n} (\mathbf{x}_i - \tilde{\mathbf{x}}_i)(g_i - \tilde{g}_i) = o_p(h^{-b}),\ \ E\left[\sum_{i=1}^{n} (g_i - \tilde{g}_i)^2\right] = o(h^{-b}),\ \textit{for any}\ b > 1.$$

$$(2) \sum_{j=1}^{n} (\mathbf{x}_j - \tilde{\mathbf{x}}_j)(\epsilon_j - \tilde{\epsilon}_j) = \sum_{j=1}^{n} \epsilon_j \{\mathbf{x}_j - E(\mathbf{x}|z_j)\} + R_n,\ ER_n^2 = o(h^{-b}),\ \textit{for any}\ b > 1.$$

$$(3) \Sigma_n := \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \tilde{\mathbf{x}}_i)(\mathbf{x}_i - \tilde{\mathbf{x}}_i)^\top \xrightarrow{p} \Phi_{\alpha,k},\ (4) n^{1/2}(\hat{\beta}_\alpha - \beta_\alpha) \xrightarrow{d} N(0, \Phi_{\alpha,k}^{-1}),\ n \to \infty.$$

Therefore, $\delta_n = \hat{\beta} - \beta = O_p(n^{-1/2})$.

Proof Theorem 1.1 As $y_i - \hat{y}^{\setminus i}(z_i) = (1-w_{ii})^{-1}(y_i - \tilde{y}_i - (\mathbf{x}_i - \tilde{\mathbf{x}}_i)^\top \hat{\beta})$, we have

$$\begin{aligned}
CV_s(\mathcal{M}_\alpha, z) &= \frac{1}{n}\sum_{i=1}^n \frac{1}{(1-w_{ii})^2}\{\epsilon_i - \tilde{\epsilon}_i + (\mathbf{x}_i - \tilde{\mathbf{x}}_i)^\top \delta_n + g_i - \tilde{g}_i\}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\epsilon_i^2}{(1-w_{ii})^2} + \frac{1}{n}\sum_{i=1}^n \frac{\tilde{\epsilon}_i^2}{(1-w_{ii})^2} + \frac{1}{n}\sum_{i=1}^n \frac{\delta_n{}^\top (\mathbf{x}_i - \tilde{\mathbf{x}}_i)(\mathbf{x}_i - \tilde{\mathbf{x}}_i)^\top \delta_n}{(1-w_{ii})^2} \\
&\quad -\frac{2}{n}\sum_{i=1}^n (\frac{\epsilon_i \tilde{\epsilon}_i}{(1-w_{ii})^2} + \frac{1}{n}\sum_{i=1}^n \frac{(g_i - \tilde{g}_i)^2}{(1-w_{ii})^2} + \frac{2}{n}\sum_{i=1}^n \frac{(\epsilon_i - \tilde{\epsilon}_i)(\mathbf{x}_i - \tilde{\mathbf{x}}_i)^\top \delta_n}{(1-w_{ii})^2} \\
&\quad +\frac{2}{n}\sum_{i=1}^n \frac{(\epsilon_i - \tilde{\epsilon}_i)(g_i - \tilde{g}_i)}{(1-w_{ii})^2} + \frac{2}{n}\sum_{i=1}^n \frac{(\mathbf{x}_i - \tilde{\mathbf{x}}_i)^\top \delta_n (g_i - \tilde{g}_i)}{(1-w_{ii})^2} \\
&:= T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8.
\end{aligned}$$

By (A.1), simple algebra leads to

$$\begin{aligned}
T_1 &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 + \frac{2}{n}\sum_{i=1}^n w_{ii}\epsilon_i^2 + o_p(\frac{1}{nh}) \\
T_2 &= \frac{1}{n}\sum_{i \neq j} w_{ij}^2 \epsilon_j^2 + \frac{1}{n}\sum_{j \neq k} \epsilon_j \epsilon_k \left(\sum_{i=1}^n w_{ij} w_{ik}\right) + o_p(\frac{1}{nh}) \\
T_4 &= \frac{2}{n}\sum_{i=1}^n w_{ii}\epsilon_i^2 + \frac{1}{n}\sum_{i \neq j} w_{ij}\epsilon_i \epsilon_j + o_p(\frac{1}{nh}), \quad T_5 = \frac{1}{n}\sum_{i=1}^n (g_i - \tilde{g}_i)^2 + o_p(\frac{1}{nh}).
\end{aligned}$$

The Theorem thus follows from Lemma A.3 and A.4.

Lemma A.3 $T_3 = O_p(n^{-1})$, $T_6 = o_p(n^{-3/2}h^{-b})$, $T_8 = o_p(n^{-3/2}h^{-b})$ *for any* $b > 3/2$.

Proof: Let $T_{81} = \sum_{i=1}^n w_i(\mathbf{x}_i - \tilde{\mathbf{x}}_i)(g_i - \tilde{g}_i)$. By (A.1) and Cauchy-Schwartz inequality,

$$
\begin{aligned}
Cov\Big\{ \textstyle\sum_{i=1}^{n} w_i\epsilon_i(\mathbf{x}_i-\tilde{\mathbf{x}}_i)\Big\} \quad &= \sigma^2 O(\frac{1}{n^2h^2})\sum_{l=1}^{n}(\mathbf{x}_i-\tilde{\mathbf{x}}_i)(\mathbf{x}_i-\tilde{\mathbf{x}}_i)^\top = O(\frac{1}{nh^2}) \\
Cov\Big\{ \sum_{i=1}^{n} w_i\tilde{\epsilon}_i(\mathbf{x}_i-\tilde{\mathbf{x}}_i)\Big\} \quad &= \sigma^2 \sum_{j=1}^{n}\sum_{i=1}^{n} w_iw_{ij}(\mathbf{x}_i-\tilde{\mathbf{x}}_i)\sum_{i=1}^{n} w_iw_{ij}(\mathbf{x}_i-\tilde{\mathbf{x}}_i)^\top \\
&\le n\sigma^2\sum_{j=1}^{n}\sum_{i=1}^{n} w_i^2w_{ij}^2(\mathbf{x}_i-\tilde{\mathbf{x}}_i)(\mathbf{x}_i-\tilde{\mathbf{x}}_i)^\top \\
&\le n\sigma^2\sum_{i=1}^{n}(\mathbf{x}_i-\tilde{\mathbf{x}}_i)(\mathbf{x}_i-\tilde{\mathbf{x}}_i)^\top w_i^3 = O(n^{-1}h^{-3}) \\
E(T_{81}T_{81}^\top) \quad &\le \sum_{i=1}^{n} w_i(\mathbf{x}_i-\tilde{\mathbf{x}}_i)(\mathbf{x}_i-\tilde{\mathbf{x}}_i)^\top\sum_{i=1}^{n} w_i(g_i-\tilde{g}_i)^2 \\
&= o(n^{-1}h^{-2-b}), \text{ for any } b>1.
\end{aligned}
$$

To prove Lemma A.4, we will repeatedly refer to the following result in [17](pp.48)

$$
\frac{1}{nh}\sum_{l=1}^{n} K_{il} = f_i + O_p(c_n), \qquad \text{uniformly in } i \text{ where } c_n = h^2 + \left(\frac{\log n}{nh}\right)^{1/2}. \qquad \text{(A.2)}
$$

Lemma A.4

$$
(1)\ \sum_{i,j}^{n} w_{ij}^2 = O_p(\frac{1}{h}),\ (2)\sum_{j\neq k}\epsilon_j\epsilon_k\sum_{i=1}^{n} w_{ij}w_{ik} = o_p(\frac{1}{h}),\ (3)\sum_{i=1}^{n} w_{ii}\epsilon_i^2 \xrightarrow{p} \sigma^2h^{-1}c_k,
$$

$$
(4)\frac{1}{n}\sum_{i\neq j} w_{ij}^2\epsilon_j^2 = \frac{\sigma^2R_K}{nh} + o_p(\frac{1}{nh}),\ \sum_{i\neq j} w_{ij}\epsilon_i\epsilon_j = o_p(h^{-1}),\ (5)\ T_7 = o_p(h^{-1}).
$$

Proof

(1) $\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{n} w_{ij}^2 \le \sum\limits_{i=1}^{n} w_{ii} = O_p(h^{-1})$.

(2) $E\Big(\sum\limits_{j\neq k}\epsilon_j\epsilon_k\sum_{i=1}^{n} w_{ij}w_{ik}\Big)^2 = O(\frac{1}{n^4h^4})E\Big(\sum\limits_{k\neq l}\sum\limits_{i,j} K_{ik}K_{il}K_{jk}K_{jl}\Big) = O(h^{-1})$.

(3) The first equation can be verified by law of large numbers and the second by (1).

(4) As $n^{-1}\sum_{i\neq j} w_{ij}^2\epsilon_j^2 = U\{1+O_p(c_n)\}$, where $U = n^{-3}h^{-2}\sum\limits_{i<j}(\epsilon_j^2K_{ji}^2f_i^{-2} + \epsilon_i^2K_{ji}^2f_j^{-2})$, it suffices to show that $U = (nh)^{-1}\sigma^2R_K + o_p(n^{-1}h^{-1})$. Define

$$H(e_i,e_j) = \epsilon_j^2f_i^{-2}K_{ij}^2 + \epsilon_i^2f_j^{-2}K_{ij}^2,\ H_1(e_i) = E\{H(e_i,e_j)|e_i\},\ H_0 = E\{H(e_i,e_j)\},$$

$$C_1(z_i) = f_i^{-2}E(K_{ij}^2|e_j),\quad C_2(z_i) = E(f_j^{-2}K_{ij}^2|e_j),\quad C_0 = E(C_1(z)) = E\{C_2(z)\}.$$

Mimicking Hoeffding's projection([40]), we have

$$\begin{aligned}U = &\frac{n-1}{2n^2h^2}H_0 + \frac{n-1}{n^3h^2}\sum_{i=1}^n\{H_1(e_i)-H_0\} + \frac{1}{2n^3h^2}\sum_{i<j}\{H(e_i,e_j)-H_1(e_i)-H_1(e_j)+H_0\}\\ :=&U_1+U_2+U_3,\end{aligned}$$

and $U_1 = (nh)^{-1}\sigma^2R_K + O(n^{-1}h)$ since $H_0 = 2h\sigma^2R_K + O(h^3)$. Note that

$$\begin{aligned}&H_1(e_i)-H_0 = \sigma^2[C_1(z_i)-C_0] + (\epsilon_i^2-\sigma^2)C_2(z_i) + \sigma^2[C_2(z_i)-C_0],\\ &EU_2^2 \le 3n^{-4}h^{-4}\sigma^4E\Big[\sum_{i=1}^n\{C_2(z_i)-C_0\}^2 + \{C_1(z_i)-C_0\}^2 + \frac{\mu_4}{\sigma^4}\{C_2(z_i)\}^2\Big]\end{aligned}$$

where $\mu_4 = E(\epsilon_i^2-\sigma^2)^2$. Therefore, $U_2 = O_p(n^{-3/2}h^{-1})$, since $E[C_k(z_i)]^2 = O(h^2),\ k=1,2$, by $(A1)$ and $(A4)$. Similarly, $U_3 = o_p(n^{-1}h^{-1})$, since

$$H(e_i,e_j)-H_1(e_i)-H_1(e_j)+H_0 = H(e_i,e_j)-H_0-(H_1(e_i)-H_0)-(H_1(e_j)-H_0).$$

and $H(e_i,e_j)-H_0$ can be argued much the same way as $H_1(e_i)-H_0$.

(5) Let $T_{71} = \sum\limits_{i=1}^n \epsilon_i(g_i-\tilde g_i)(1-w_{ii})^{-2}$, $T_{72} = \sum\limits_{i=1}^n \tilde\epsilon_i(g_i-\tilde g_i)(1-w_{ii})^{-2}$. Then by Lemma A.2(1), $T_{71} = o_p(h^{-1})$. Next, we have

$$\begin{aligned}\Big|\sum_{i=1}^n\tilde\epsilon_i(g_i-\tilde g_i)\Big| \le &\frac{c_n}{nh}\sum_{i=1}^n|(g_i-\tilde g_i)\sum_{j=1}^nK_{ij}\epsilon_j| + \frac{1}{nh}|\sum_{i=1}^nf_i^{-1}(g_i-\tilde g_i)\sum_{j=1}^nK_{ij}\epsilon_j|\\ =& S_1+S_2.\end{aligned}\tag{A.3}$$

By Cauchy-Schwarz inequality and Lemma A.2(1),

$$ES_1^2 \leq \frac{\sigma^2 c_n^2}{n^2h^2} E\Big\{\sum_{i=1}^n (g_i - \tilde{g}_i)^2\Big\} E\Big\{\sum_{i=1}^n \sum_{j=1}^n K_{ij}^2\Big\} = o(c_n^2 h^{-b}), \text{ for any } b > 2,$$

Therefore, $S_1 = o(h^{-1})$.

$$S_2 \leq \Big|\sum_{i,j,k} \frac{K_{ij}K_{ik}\epsilon_j(g_i - g_k)}{n^2h^2 f_i^2}\Big| + O_p(\frac{c_n}{n^2h^2}) \sum_{i=1}^n f_i^{-2} \Big|\sum_{j,k} K_{ij}\epsilon_j K_{ik}(g_i - g_k)\Big|,$$

where the terms inside $|.|$ can be dealt with using Lemma A.1. Take the first term for example. Define

$$\begin{aligned} H(e_i, e_j, e_k) = & \frac{K_{ij}K_{ik}}{f^2(z_i)}\epsilon_j(g_i - g_k) + \frac{K_{kj}K_{ki}}{f^2(z_k)}\epsilon_j(g_k - g_i) + \frac{K_{ji}K_{jk}}{f^2(z_j)}\epsilon_i(g_j - g_k) \\ & + \frac{K_{ki}K_{kj}}{f^2(z_k)}\epsilon_i(g_k - g_j) + \frac{K_{ji}K_{jk}}{f^2(z_j)}\epsilon_k(g_j - g_i) + \frac{K_{ij}K_{ik}}{f^2(z_i)}\epsilon_k(g_i - g_j). \end{aligned}$$

Hence $EH(e_i, e_j, e_k) = 0$ and

$$H_1(e_i) = E[H(e_i, e_j, e_k)|e_i] = \epsilon_i E_{j,k}\Big\{K_{ji}K_{jk}f_j^{-2}(g_j - g_k) + K_{ki}K_{kj}f_k^{-2}(g_k - g_j)\Big\}.$$

Let $s = z_{ij}/h,\ t = z_{jk}/h$,

$$\begin{aligned} & \int K_{ij}K_{jk}(g_j - g_k)f_j^{-1}f_k dz_j dz_k \\ & = h^2 \int K(s)K(t)\Big\{g(z_i + hs) - g(z_i + hs - ht)\Big\}f^{-1}(z_i + hs)f(z_i + hs - ht)dsdt \\ & = h^2 \int K(s)K(t)htg'(z_i)\Big\{1 - htf'(z_i)f_i^{-1} + o(h)\Big\}, \end{aligned}$$

where the last equality holds since $g(t) = \beta_k t$. Therefore, by (A4), $H_1(e_i) = O_p(h^4)$ and $EH_1^2(e_i) = ch^8 + o(h^8)$ with $c > 0$. By Lemma A.1, $n^{-5/2}\sum_{i,j,k} H(e_i, e_j, e_k) = O_p(h^4)$. Therefore, if $nh^6 \to 0$, we have

$$\sum_{i,j,k} \frac{K_{ij}K_{ik}\epsilon_j(g_i - g_k)}{n^2h^2 f^2(z_i)} = \frac{1}{6}n^{-2}h^{-2}\sum_{i,j,k} H(e_i, e_j, e_k) = O_p(n^{1/2}h^2) = o_p(h^{-1}).$$

Appendix B

Conditions and Proofs for Chapter 2

First we introduce some notation. Let $\gamma_\alpha(.|\theta)$ and $\gamma_0(.)$ be the density functions of $X_\alpha^\top\theta_\alpha$ and $X^\top\theta^0$ respectively. Let $\mathcal{U}_\alpha = \{X_\alpha^\top\theta_\alpha^0 : X \in A\}$, $\mathcal{U}_\alpha = \{X_\alpha^\top\theta_\alpha^0 : X \in A\}$, $\mathcal{D}_\alpha = \{x_\alpha : x \in A\}$, $A_d = \{x_d : (x_1, \cdots, x_d, \cdots, x_p) \in A\}$, with A defined in Section 2, and $K_1 = \int t^2 K(t)dt$, $K_2 = \int K^2(t)dt$. For any $\alpha \supset \alpha_0$, let $\mu_\alpha(x|\theta) = E(X_\alpha|X_\alpha^\top\theta_\alpha = x_\alpha^\top\theta_\alpha)$, $v_\alpha(x|\theta) = \mu_\alpha(x|\theta) - x_\alpha$, $W_\alpha = \int_A v_\alpha(x|\theta)v_\alpha(x|\theta)^\top g'(x^\top\theta^0)^2 f(x)dx$, and

$$U_{\alpha,j} = W_\alpha^{\frac{1}{2}+} g'(X_j^\top\theta^0)v_\alpha(X_j|\theta^0).$$

Let $\Theta_{n,\alpha} = \{\theta_\alpha : \|\theta\| = 1, \|\theta - \theta^0\| \le rn^{-1/2}\}$, $\mathcal{H}_n = \{h : r_1 n^{-1/5} \le h \le r_2 n^{-1/5}\}$ for some $r > 0$ and $0 < r_1 < r_2 < \infty$. Denote $\Theta_{n,\{1,\cdots,p\}}$ by Θ_n.

We impose the following regularity conditions to prove the theorems.

(A1) X has a compact support in $\mathcal{R}^p$ and for any $\alpha \supseteq \alpha_0$, $\inf_{x\in A,\theta\in\Theta_{n,\alpha}} \gamma_\alpha(x^\top\theta|\theta) > 0$.

(A2) The link function $g(.)$ has bounded third-order derivatives on $\mathcal{U}$.

(A3) The function K is a symmetric density function with a compact support. Assume that $K_1 = 1$ and the Fourier transform of $K(t)$ is absolutely integrable.

(A4) We have $E(\epsilon_i|X_i) = 0$ and $E(\epsilon_i^2|X_i) = \sigma^2$.

(A5) For any α, $\sup\limits_{m\to\infty}\sup\limits_{s}\|m^{-1}\sum\limits_{j\in s} U_{\alpha,j}U_{\alpha,j}^\top - I_{d_\alpha} + \theta_\alpha^0\theta_\alpha^{0^\top}\| = o_p(1)$, and

$$\hat{\theta}_\alpha^{\backslash s} - \theta_\alpha^0 = {n_c}^{-1}W_\alpha^+\sum_{j\notin s} g'(X_{j,\alpha}^\top\theta_\alpha^0)v_\alpha^\top(X_j|\theta^0)\epsilon_j + \delta_n^{\backslash s}, \tag{B.1}$$

where I_{d_α} is the identity matrix and $\delta_n^{\backslash s} = o_p(n^{-1/2})$ uniformly for all s.

(A6) For any $\alpha \subset \alpha_0$, $g_\alpha(v|\theta) = E(Y|X_\alpha^\top\theta = v^\top\theta)$ has bounded first-order derivative with respect to $\theta \in \Theta_{n,\alpha}$; $\sigma_\alpha^2(\theta) := E\left\{g_\alpha(X_\alpha|\theta) - Y\right\}^2$ and $\inf\limits_{\theta\in\Theta_{n,\alpha}} \sigma_\alpha^2(\theta) > \sigma^2$.

(A7) Suppose $\alpha \cup d \supseteq \alpha_0$. For $\mathbf{x}_d$ continuous, the joint density function of $(X_\alpha^\top\theta, \mathbf{x}_d)$, $f_{X_\alpha^\top\theta,\mathbf{x}_d}(u^\top\theta, v)$, is uniformly bounded away from zero for $\theta \in \Theta_{n,\alpha}$, $u \in \mathcal{D}_\alpha$ and $v \in A_d$; For $\mathbf{x}_d$ discrete, the conditional density function of $X_\alpha^\top\theta$ given $x_d = v$, $f_{X_\alpha^\top\theta|x_d=v}(.)$ satisfies $\inf\limits_{u\in\mathcal{D}_\alpha,\theta\in\Theta_{n,\alpha}} f_{X_\alpha^\top\theta_\alpha^0|x_d=v}(u^\top\theta) > 0$.

Assumptions (A1)-(A4) are required for the consistency of estimations; see [34, 91] . For (A5), while [91] proved (B.1) with $\delta_n^{\backslash s} = o_p(n^{-1/2})$ for any given s, the uniform convergence rate here is necessary to guarantee the validity of leaving-m-out crossvalidation and parallels the balanced block design assumption in linear regression; see [94]. The requirement on the Fourier transform of $K(t)$ in (A3) is to ensure the difference between the MAVE estimator $\hat{\theta}$ and θ_α^0 admit the form in (B.1). Many kernel functions meet this demand, such as the triweight kernel. The Gaussian kernel is also permissible at the expense of a longer proof. (A6) is a common assumption if the optimal model exists and is unique; see [89]. (A7) is used to ensure the denominators of kernel smoothers are bounded away from zero.

Proof Theorem 2.1 We consider two cases with $m = 1$ or $m > 1$.

1. [34] proved that, for any $\alpha \supset \alpha_0$, $HCV_\alpha(\theta, h)$ defined in (4) can be written as $HCV_\alpha(\theta, h) = \tilde{S}_\alpha(\theta) + H$, where H contains terms either of higher order than $\tilde{S}_\alpha(\theta)$ or independent of θ (thus model α) and

   $$\tilde{S}_\alpha(\theta) = \sum_{i=1}^{n} e_i^2 - Z^\top Z + n(W_\alpha^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z)^\top (W_\alpha^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z) + o_p(1),$$

   where $Z = n^{-1/2}\sigma^{-1}W_\alpha^{-1/2}V_\alpha$, which is asymptotically $N_{d_\alpha}(0, I)$. Therefore, the dominating term in the deviance of HCV_α from HCV_{α_0} is given by $Z_{\alpha_0}^\top Z_{\alpha_0} - Z_\alpha^\top Z_\alpha$, which is asymptotically $\chi^2(d - d_0)$. The proof of is thus completed.

2. Let $N = m\binom{n}{m}$, $\sum\limits_{i,s} = \sum\limits_{s}\sum\limits_{i\in s}$. For any $\alpha \supset \alpha_0$, mimicking the steps in [34], we have

   $$HCV_\alpha^m(\theta, h) = \tilde{S}_\alpha(\theta) + \frac{1}{N}\sum_{i,s}\{D_i^{\backslash s2} + \Delta_i^{\backslash s2} + 2(D_i^{\backslash s}\Delta_i^{\backslash s} + \Delta_i^{\backslash s}\delta_i + D_i^{\backslash s}\delta_i - D_i^{\backslash s}\epsilon_i - \Delta_i^{\backslash s}\epsilon_i)\}$$

   where $\tilde{S}_\alpha(\theta) = \frac{1}{N}\sum\limits_{i,s}\{Y_i - g(X_{i,\alpha}^\top\theta|\theta)\}^2$, $D_i^{\backslash s} = \hat{g}_\alpha^{\backslash s}(X_i^\top\theta^0|\theta^0) - g(X_i^\top\theta^0)$,

   $\delta_i = g(X_{i,\alpha}^\top\theta|\theta) - g(X_i^\top\theta^0)$, $\Delta_i^{\backslash s} = \hat{g}_\alpha^{\backslash s}(X_{i,\alpha}^\top\theta|\theta) - g(X_{i,\alpha}^\top\theta|\theta) - \{\hat{g}_\alpha^{\backslash s}(X_i^\top\theta^0|\theta) - g(X_i^\top\theta^0)\}$.

   In outline, our argument runs as follows. We show in step 1 that with $X-$probability 1, and for all $\xi > 0$, $N^{-1}\sum_{i,s}\Delta_i^{\backslash s2} = O_p(n^{-7/5+\xi})$. As $N^{-1}\sum_{i,s}E(D_i^{\backslash s2}) = O(n^{-4/5})$, $N^{-1}\sum_{i,s}D_i^{\backslash s2} = O(n^{-4/5})$. By Taylor expansion at θ^0, it follows that

$\delta_i = O(n^{-1/2})$ uniformly in i, and $N^{-1}\sum_{i,s}\delta_i^{\backslash s2} = O(n^{-1})$ and

$$\begin{aligned} & \frac{1}{N}\sum_{i,s}\Delta_i^{\backslash s2} + \frac{2}{N}\sum_{i,s}(D_i^{\backslash s}\Delta_i^{\backslash s} + \Delta_i^{\backslash s}\delta_i) \\ \leq\ & \frac{1}{N}\sum_{i,s}\Delta_i^{\backslash s2} + \frac{2}{N}\Big(\sum_{i,s}\Delta_i^{\backslash s2}\Big)^{1/2}\Big\{(\sum_{i,s}D_i^{\backslash s})^{1/2} + (\sum_{i,s}\delta_i)^{1/2}\Big\} \\ =\ & O_p\Big\{n^{-7/5+\xi} + (n^{-7/5+\xi}n^{-4/5})^{1/2}\Big\} = o_p(n^{-1}). \end{aligned}$$

We will prove in step 2 and 3 that $N^{-1}\sum_{i,s}D_i^{\backslash s}\delta_i = O(n^{-13/10+\xi})$, and in step 4 $N^{-1}\sum_{i,s}\Delta_i^{\backslash s}\epsilon_i = O_p(n^{-11/10+\xi})$. Putting all together, we will have proved that $HCV_\alpha(\theta, h) = \tilde{S}_\alpha(\theta) + N^{-1}\sum_{i,s}(D_i^{\backslash s2} - 2D_i^{\backslash s}\epsilon_i) + o_p(n^{-1})$. As $D_i^{\backslash s}$ is independent of α, if we can prove $\tilde{S}_\alpha < \tilde{S}_{\alpha_0}$, for any $\alpha \supset \alpha_0$, this theorem will be established.

Step 1. $N^{-1}\sum_{i,s}\Delta_i^{\backslash s2} = O_p(n^{-7/5+\xi}),$ with $X - probability\ 1,\ \forall\xi > 0.$

For $s = \{i\}$, [34] proved that $E(\Delta_i^{\backslash s}) = O(n^{-7/10+\xi})$ and $Var(\Delta_i^{\backslash s}) = O(n^{-2}h^{-3})$ uniformly in i. Since $m/n \to m \in [0, 1)$, the same result holds for $\#s = m$.

Step 2. $|N^{-1}\sum_{i,s}E(D_i^{\backslash s})\delta_i| = O(n^{-13/10+\xi}),$ *with* $X - probability\ 1, \forall\xi > 0.$

For any bounded X, by Taylor expansion, we have

$$g(X^\top\theta^0) = g(X^\top\theta) - \eta(\theta_{00}^\top X)g'(X^\top\theta^0) + O(n^{-1}) \tag{B.2}$$

$$g(X_\alpha^\top\theta|\theta) = g(X_\alpha^\top\theta) - \eta\{\theta_{00}^\top\mu_\alpha(X_\alpha|\theta)\}g'(X^\top\theta^0) + O(n^{-1}). \tag{B.3}$$

Note that $\mu_\alpha(X_\alpha|\theta) - \mu_\alpha(X_\alpha|\theta^0) = O(n^{-1/2})$. Therefore

$$\delta_i = \eta\theta_{00}^\top\{X_i - \mu_\alpha(X_{i,\alpha}|\theta^0)\}g'(X^\top\theta^0) + O(n^{-1}),$$

$$E(\mathcal{D}_i^{\backslash s}) = b_i^{\backslash s}(\theta^0)c_i^{\backslash s}(\theta^0)^{-1} = b_i^{\backslash s}(\theta^0)\gamma_0(X_i)^{-1} + O(h^4n^\xi) = O(h^2n^\xi),$$

uniformly in i, where $c_i^{\backslash s}(\theta^0) = \frac{1}{(n-m)h} \sum_{j \notin s} K_h\{(X_j - X_i)^\top \theta^0\} = \gamma_0(X_i) + O(h^2 n^\xi)$,

$$b_i^{\backslash s}(\theta^0) = \frac{1}{(n-m)h} \sum_{j \notin s} \{g(X_{j,\alpha}^\top \theta^0) - g(X_{i,\alpha}^\top \theta^0)\} K_h\{(X_j - X_i)^\top \theta^0\} = O(h^2 n^\xi),$$

uniformly in i. Hence, $n^{-1} \sum_{i,s} E(D_i^{\backslash s})\delta_i = -n^{-1}\eta t + O(n^{-13/10+\xi})$, with

$$t = \sum_{i,s} \theta_{00}^\top \{X_{i,\alpha} - \mu(X_{i,\alpha}|\theta^0)\} g'(X_i^\top \theta^0) b_i^{\backslash s}(\theta^0) \gamma_0(X_i)^{-1},$$

the observed value of $T = n_c^{-1} h^{-1} \binom{n-2}{m-1} \theta_{00}^\top \sum_{j \neq i} A(X_i, X_j)$, where

$$A(X_i, X_j) = \{X_{i,\alpha} - \mu(X_{i,\alpha}|\theta^0)\} g'(X_i^\top \theta^0) \gamma_0(X_i)^{-1} \{g(X_j^\top \theta^0) - g(X_i^\top \theta^0)\} K_h(X_{ij}^\top \theta^0).$$

Again, through similar arguments in [34], we have $T = \binom{n-2}{m-1} O(n^{1/2+\xi} h^2)$. recall that $N = m\binom{n}{m}$. The desired result thus follows.

Step 3. With $X - probability$ 1, $\forall \xi > 0, Var\Big(N^{-1} \sum_{i,s} D_i^{\backslash s} \delta_i\Big) = O(n^{-14/5+\xi})$.

Simple algebra gives

$$Var\Big(\sum_{i,s} D_i^{\backslash s} \delta_i\Big) = \frac{1}{(n-m)^2 h^2} \sum_j u_j^2 \sigma_j^2, \ u_j = \sum_{j \notin s} \sum_{i \in s} \delta_i c_i^{\backslash s}(\theta^0)^{-1} K_h\{X_{ij}^\top \theta^0\}$$

Similarly to that in step 2, we can prove that with $X-$probability 1 and for all $\xi > 0$, $u_j = -\eta v_j + k\binom{n}{m} O(n^{-1/2+\xi} h^2)$ uniformly in j, where

$$v_j = \sum_{j \notin s} \sum_{i \in s} \theta_{00}^\top \{X_{i,\alpha} - \mu_\alpha(X_{i,\alpha}|\theta^0)\} g'(X_i^\top \theta^0) \gamma_0(X_i)^{-1} K_h\{X_{ij}^\top \theta^0\}.$$

Therefore, $Var\Big(\sum_{i,s} D_i^{\backslash s} \delta_i\Big) \leq 2(n-m)^{-2} h^{-2} \eta^2 \sum_j v_j^2 \sigma_j^2 + k^2 \binom{n-1}{m}^2 O(n^{2\xi-2} h^2)$. Now v_j equals the observed value of $V_j = \binom{n-2}{m-1} \sum_{i \neq j} B(X_i, X_j)$, where

$$B(X_i, X_j) = \theta_{00}^\top \{X_{i,\alpha} - \mu_\alpha(X_{i,\alpha}|\theta^0)\} g'(X_i^\top \theta^0) \gamma_0(X_i)^{-1} K_h\{X_{ij}^\top \theta^0\}.$$

By similar arguments used for $A(X_i, X_j)$ in step 2, we have $X-$probability 1,

$$\begin{aligned}Var\Big(\sum_{i,s} D_i^{\setminus s}\delta_i\Big) =& O\Big\{n^{-2}h^{-2}\eta^2\binom{n-2}{m-1}^2 n^{2+\xi}h + m^2\binom{n-1}{m}^2 O(n^{-1+2\xi}h^4)\Big\}\\ =& O\Big(n^{\xi-1}h^{-1}\binom{n-2}{m-1}^2 + \binom{n-1}{m}^2 O(n^{1+2\xi}h^4)\Big), \forall\xi > 0\end{aligned}$$

Therefore, $Var\Big(\frac{1}{n}\sum\limits_{i,s} D_i^{\setminus s}\delta_i\Big) = O(n^{\xi-3}h^{-1})$ which completes the proof.

Step 4. With $X-$probability 1, $E\Big(n^{-1}\sum\limits_{i,s}\Delta_i^{\setminus s}\epsilon_i\Big)^2 = O_p(n^{-11/5+\xi}), \forall\xi > 0$

Since $E(\Delta_i^{\setminus s}) = O(n^{-7/10+\xi})$ uniformly in i (see step 1), $E\Big\{n^{-1}\sum\limits_{i,s} E(\Delta_i^{\setminus s})\epsilon_i\Big\}^2 = n^{-1}O(n^{-7/5+\xi})$. For any two subsets s_1, s_2 of $\{1, \cdots, p\}$, define $S_{ij}^{s_1,s_2} = E\Big\{(\Delta_i^{\setminus s_1} - E\Delta_i^{\setminus s_1})\epsilon_i(\Delta_j^{\setminus s_2} - E\Delta_j^{\setminus s_2})\epsilon_j\Big\}$. Therefore,

$$S_{ii}^{s_1,s_1} = Var(\Delta_i^{\setminus s_1})\sigma_i^2, \quad S_{ii}^{s_1,s_2} \leq 2\Big\{Var(\Delta_i^{\setminus s_1}) + Var(\Delta_i^{\setminus s_2})\Big\}\sigma_i^2.$$

For fixed $s_1 \ni i \neq j \in s_2$, if $i \in s_2$, or $j \in s_1$, $S_{ij}^{s_1,s_2} = 0$; otherwise,

$$\begin{aligned}S_{ij}^{s_1,s_2} =& \sigma_i^2\sigma_j^2\Big\{\frac{K_h(X_{ij}^\top\theta^0)}{\sum\limits_{k\notin s_1} K_h((X_k - X_i)^\top\theta^0)} - \frac{K_h((X_{j,\alpha} - X_{i,\alpha})^\top\theta)}{\sum\limits_{k\notin s_1} K_h((X_{k,\alpha} - X_{i,\alpha})^\top\theta)}\Big\}\\ &\Big\{\frac{K_h(X_{ij}^\top\theta^0)}{\sum\limits_{k\notin s_2} K_h((X_k - X_i)^\top\theta^0)} - \frac{K_h((X_{j,\alpha} - X_{i,\alpha})^\top\theta)}{\sum\limits_{k\notin s_2} K_h((X_{k,\alpha} - X_{i,\alpha})^\top\theta)}\Big\}.\end{aligned}$$

Note that

$$\begin{aligned}\Delta_i^{\setminus s_1} - E\Delta_i^{\setminus s_1} =& \sum_{j\notin s_1}\epsilon_j\Big\{\frac{K_h(X_{ij}^\top\theta^0)}{\sum\limits_{j\notin s_1} K_h(X_{ij}^\top\theta^0)} - \frac{K_h((X_{j,\alpha} - X_{i,\alpha})^\top\theta)}{\sum\limits_{j\notin s_1} K_h((X_{j,\alpha} - X_{i,\alpha})^\top\theta)}\Big\},\\ E\Big\{\frac{1}{N}\sum_{i,s}(\Delta_i^{\setminus s} - E\Delta_i^{\setminus s})\epsilon_i\Big\}^2 =& \frac{1}{N^2}\Big\{\sum_{i,s} S_{ii}^{s,s} + \sum_{i\in s_1\cap s_2}^{s1\neq s2} S_{ii}^{s_1,s_2} + \sum_{s_1\ni i\neq j\in s_2}^{i\notin s_2, j\notin s_1} S_{ij}^{s_1,s_2}\Big\}\\ =& O(n^{-4}h^{-3} + n^{-3}h^{-4}) = O(n^{-3}h^{-4}).\end{aligned}$$

The desired results thus follows.

Step 5 By $(B.2)$ and $(B.3)$, we have

$$g(X_i^\top \theta^0) - g(X_{i,\alpha}^\top \theta|\theta) = \eta\theta_{00}^\top\{\mu_\alpha(X_{i,\alpha}|\theta^0) - X_{i,\alpha}\}g'(X_i^\top \theta^0) + O(n^{-1})$$

Define $W = \sum_{i,s}\{X_{i,\alpha} - \mu_\alpha(X_{i,\alpha}|\theta^0)\}\{X_{i,\alpha} - \mu_\alpha(X_{i,\alpha}|\theta^0)\}^\top g'(X_i^\top\theta^0)^2$, then

$$\begin{aligned}
\tilde{S}_\alpha(\theta) &= \frac{1}{N}\sum_{i=1}^n\{\epsilon_i + g(X_i^\top\theta^0) - g(X_{i,\alpha}^\top\theta|\theta)\}^2 \\
&= \frac{1}{N}\sum_{i=1}^n \epsilon_i^2 - \frac{2}{n}\eta\theta_{00}^\top V_\alpha + \frac{1}{n}\theta_{00}^\top W\theta_{00} + o(n^{-1}) \\
&= \frac{1}{N}\Big\{\sum_i e_i^2 - Z^\top Z + n(W_{\alpha_0}^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z)^\top(W_\alpha^{1/2}\eta\theta_{00} - n^{-1/2}\sigma Z) + o_p(1)\Big\}
\end{aligned}$$

which parallels the result in HHI, thus implies the same conclusion.

Before proceeding to the proof of Theorem 2.2, we introduce the following lemma of [91]

Lemma B.1 *[Basic results for kernel smoothing] If $E(Z|\theta^\top X = \theta^\top x) = m_\theta(x)$ has bounded third derivatives, and $\mu_d^K = \int K(v)v^d dv$, then*

$$\frac{1}{nh}\sum_{i=1}^n K_h(\theta^\top(X_i - x))[\theta^\top(X_i - x)/h]^d Z_i = f_\theta(x)m_\theta(x)\mu_d^K + \{f_\theta(x)m_\theta(x)\}'\,\mu_{d+1}^K h + O(\tau_n),$$

uniformly for $(\theta, x) \in \Theta_n \bigotimes A^{bh}$, where $\tau_n = h^2 + (\log n/nh)^{1/2}$.

Proof Theorem 2.2 If $\alpha \supset \alpha_0$, and $h \in \mathcal{H}_n$, by Lemma 9 in [91], the local linear estimator $\hat{g}_\alpha(u|\theta)$ based on $\{X_i, Y_i\}_{i=1}^n$ has the following expression

$$\begin{aligned}
\hat{g}_\alpha(u|\theta) &= g(u^\top\theta_\alpha^0) - g'(u^\top\theta_\alpha^0)(\theta_\alpha^0 - \theta)^\top v_\alpha(u|\theta_\alpha^0) + \frac{1}{2}g''(u^\top\theta_\alpha^0)h^2 \\
&\quad + n^{-1}\gamma_\alpha^{-1}(u^\top\theta|\theta)\sum_{i=1}^n K_h\{(X_{i,\alpha} - u)^\top\theta\}\epsilon_i + r_n(u|\theta), \qquad \text{(B.4)}
\end{aligned}$$

where $r_n(u|\theta) = o_p(n^{-1/2})$ uniformly for $u \in \mathcal{D}_\alpha$ and $\theta \in \Theta_{n,\alpha}$. The above equation continues to hold if we consider the leave-m-out estimator, i.e.

$$\begin{aligned}\hat{g}_\alpha^{\backslash s}(u|\theta) &= g(u^\top\theta_\alpha^0) - g'(u^\top\theta_\alpha^0)(\theta_\alpha^0 - \theta)^\top v_\alpha(u|\theta_\alpha^0) + \frac{1}{2}g''(u)h^2 \\ &\quad + n_c^{-1}\gamma_\alpha^{-1}(u^\top\theta|\theta)\sum_{j\notin s} K_h\{(X_{j,\alpha} - u)^\top\theta\}\epsilon_j + r_n^{\backslash s}(u|\theta),\end{aligned}$$

where $r_n^{\backslash s}(u|\theta) = o_p(n^{-1/2})$ uniformly for $u \in \mathcal{D}_\alpha$, $\theta \in \Theta_{n,\alpha}$ and all s. Since $\hat{\theta}_\alpha^{\backslash s} \in \Theta_{n,\alpha}$ by (B.1), we have

$$\hat{g}_\alpha^{\backslash s}(u|\theta_\alpha^0) - \hat{g}_\alpha^{\backslash s}(u|\hat{\theta}_\alpha^{\backslash s}) = g'(u^\top\theta_\alpha^0)v_\alpha(u|\theta^0)(\hat{\theta}_\alpha^{\backslash s} - \theta_\alpha^0) + R(u|\theta_\alpha^0, \hat{\theta}_\alpha^{\backslash s})$$

uniformly for all s, where

$$\begin{aligned}R(u|\theta_\alpha^0, \hat{\theta}_\alpha^{\backslash s}) = &\frac{1}{n_c}\Big\{\gamma_\alpha^{-1}(u^\top\theta_\alpha^0|\theta_\alpha^0) - \gamma_\alpha^{-1}(u^\top\hat{\theta}_\alpha^{\backslash s}|\hat{\theta}_\alpha^{\backslash s})\Big\}\sum_{j\notin s} K_h\{(X_{i,\alpha} - u)^\top\theta_\alpha^0\}\epsilon_j \\ &+ r_n^{\backslash s}(u|\theta_\alpha^0) - r_n^{\backslash s}(u|\hat{\theta}_\alpha^{\backslash s}).\end{aligned}$$

It follows from Lemma 7 in [91] that

$$\begin{aligned}&\frac{1}{n_c}\Big\{\gamma_\alpha^{-1}(u^\top\theta_\alpha^0|\theta_\alpha^0) - \gamma_\alpha^{-1}(u^\top\hat{\theta}_\alpha^{\backslash s}|\hat{\theta}_\alpha^{\backslash s})\Big\}\sum_{j\notin s} K_h\{(X_{i,\alpha} - u)^\top\theta_\alpha^0\}\epsilon_j = O_p(n^{-1/2}\tau_n), \\ &\frac{1}{n_c}\sum_{j\notin s}\Big[K_h\{(X_{i,\alpha} - u)^\top\theta_\alpha^0\} - K_h\{(X_{j,\alpha} - u)^\top\hat{\theta}_\alpha^{\backslash s}\}\Big]\epsilon_j = O_p(n^{-1/2}\tau_n).\end{aligned}$$

Therefore $R(u|\hat{\theta}_\alpha^0, \hat{\theta}_\alpha^{\backslash s}) = o_p(n^{-1/2})$ uniformly for all $u \in \mathcal{D}_\alpha$ and all s, as long as $\tau_n \to 0$. Recall that $U_{\alpha,j} = W_\alpha^{\frac{1}{2}+}g'(X_j^\top\theta^0)v_\alpha(X_j|\theta^0)$. We have

$$\begin{aligned}Y_i - \hat{g}_\alpha^{\backslash s}(X_i|\hat{\theta}_\alpha^{\backslash s}) &= Y_i - \hat{g}_\alpha^{\backslash s}(X_i|\theta_\alpha^0) + \hat{g}_\alpha^{\backslash s}(X_i|\theta_\alpha^0) - \hat{g}_\alpha^{\backslash s}(X_i|\hat{\theta}_\alpha^{\backslash s}) \\ = \quad &Y_i - \hat{g}_\alpha^{\backslash s}(X_i|\theta_\alpha^0) + n_c{}^{-1}U_{\alpha,i}^\top\sum_{j\notin s} U_{\alpha,j}\epsilon_j + R(X_{i,\alpha}|\theta_\alpha^0, \hat{\theta}_\alpha^{\backslash s}) + g_i' v_\alpha^\top(X_i|\theta_\alpha^0)\delta_n^{\backslash s}.\end{aligned}$$

$$
\begin{aligned}
CV_\alpha(m) = m^{-1}\binom{n'}{m}^{-1} &\Big[{\sum_{i,s}}'\{Y_i - \hat{g}_\alpha^{\setminus s}(X_i|\theta_\alpha^0)\}^2 + \frac{1}{n_c^2}{\sum_{i,s}}' U_{\alpha,i}^\top\{\sum_{j\notin s} U_{\alpha,j}\epsilon_j\}\{\sum_{j\notin s} U_{\alpha,j}\epsilon_j\}^\top U_{\alpha,i} \\
&+\frac{2}{n_c}{\sum_{i,s}}'\{Y_i - \hat{g}_\alpha^{\setminus s}(X_i|\theta_\alpha^0)\}U_{\alpha,i}^\top \sum_{j\notin s} U_{\alpha,j}\epsilon_j + 2{\sum_{i,s}}'\{Y_i - \hat{g}_\alpha^{\setminus s}(X_i|\theta_\alpha^0)\}R(X_{i,\alpha}|\theta_\alpha^0, \hat{\theta}_\alpha^{\setminus s}) \\
&+2{\sum_{i,s}}'\{Y_i - \hat{g}_\alpha^{\setminus s}(X_i|\theta_\alpha^0)\}g'(X_{i,\alpha}^\top\theta_\alpha^0)v_\alpha^\top(X_i|\theta^0)\delta_n^{\setminus s} + \frac{2}{n_c}{\sum_{i,s}}' R(X_{i,\alpha}|\theta_\alpha^0, \hat{\theta}_\alpha^{\setminus s})U_{\alpha,i}^\top \sum_{j\notin s} U_{\alpha,j}\epsilon_j \\
&+\frac{2}{n_c}{\sum_{i,s}}' g'(X_{i,\alpha}^\top\theta_\alpha^0)v_\alpha^\top(X_i|\theta^0)U_{\alpha,i}\sum_{j\notin s} U_{\alpha,j}^\top\epsilon_j\delta_n^{\setminus s}\Big] + o_p(\frac{1}{n}) \\
&:= RSS(m) + T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + o_p(\frac{1}{n}), \qquad \text{(B.5)}
\end{aligned}
$$

where the term $o_p(\frac{1}{n})$ is established by (A1),(A6) and the following fact

$$
\frac{1}{m\binom{n'}{m}}{\sum_{i,s}}' R^2(X_{i,\alpha}|\theta_\alpha^0, \hat{\theta}_\alpha^{\setminus s}) = o_p(\frac{1}{n}), \ \frac{1}{m\binom{n'}{m}}{\sum_{i,s}}'\{g'(X_{i,\alpha}^\top\theta_\alpha^0)v_\alpha^\top(X_i|\theta^0)\delta_n^{\setminus s}\}^2 = o_p(\frac{1}{n})
$$

$$
\Big|R(X_{i,\alpha}|\theta_\alpha^0, \hat{\theta}_\alpha^{\setminus s})g'(X_{i,\alpha}^\top\theta_\alpha^0)v_\alpha^\top(X_i|\theta^0)\delta_n^{\setminus s}\Big| \le R^2(X_{i,\alpha}|\theta_\alpha^0, \hat{\theta}_\alpha^{\setminus s}) + \{g'(X_{i,\alpha}^\top\theta_\alpha^0)v_\alpha^\top(X_i|\theta^0)\delta_n^{\setminus s}\}^2
$$

In (B.5) $T_5 = o_p(n^{-1})$ and $T_6 = o_p(n^{-1})$ can be verified by calculating the second moments. Next, we prove that

$$
T_2 = \frac{2}{mn_c}\binom{n'}{m}^{-1}{\sum_{i,s}}'\epsilon_i U_{\alpha,i}^\top \sum_{j\notin s} U_{\alpha,j}\epsilon_j + o_p(\frac{1}{n}). \qquad \text{(B.6)}
$$

Let $\tilde{\epsilon}_i^{\setminus s} \equiv n_c^{-1}\sum\limits_{j\notin s} K_h(X_{ij}^\top\theta^0)\epsilon_j\gamma_0^{-1}(X_i^\top\theta^0)$. Then, by (B.5),

$$
Y_i - \hat{g}_\alpha^{\setminus s}(X_i|\theta_\alpha^0) = \epsilon_i - \tilde{\epsilon}_i^{\setminus s} + h^2 g''(X_i^\top\theta^0)/2 + o_p(n^{-1/2}),
$$

with the $o_p(n^{-1/2})$ term independent of X_i, thus $U_{\alpha,i}$. As $U_{\alpha,i}$ is independent of ϵ_j $(i \neq j)$,

$$
E\{\sum_{i\in A}^{j\notin A} g''(X_i^\top\theta^0)U_{\alpha,i}^\top U_{\alpha,j}\epsilon_j\}^2 = O(n'(n-n')), \ E\{\sum_{i,j\in A}^{i\neq j} g''(X_i^\top\theta^0)U_{\alpha,i}^\top U_{\alpha,j}\epsilon_j\}^2 = O(n'^2), \ \text{(B.7)}
$$

$$
{\sum_{i,s}}' g''(X_i^\top\theta^0)U_{\alpha,i}^\top \sum_{j\notin s} U_{\alpha,j}\epsilon_j = \tbinom{n'-1}{m-1}\sum_{i\in A}^{j\notin A} g''(X_i^\top\theta^0)U_{\alpha,i}^\top U_{\alpha,j}\epsilon_j + \tbinom{n'-2}{m-1}\sum_{i,j\in A}^{i\neq j} g''(X_i^\top\theta^0)U_{\alpha,i}^\top U_{\alpha,j}\epsilon_j,
$$

and $n - n' = O(nh)$, thus

$$\frac{h^2}{mn_c}\binom{n'}{m}^{-1}{\sum_{i,s}}' g''(X_i^\top\theta^0)U_{\alpha,i}^\top\sum_{j\notin s}U_{\alpha,j}\epsilon_j = \frac{h^2}{mn_c}\binom{n'}{m}^{-1}\{\binom{n'-1}{m-1}O_p(nh^{\frac{1}{2}}) + \binom{n'-2}{m-1}O_p(n')\}$$
$$= O_p\{(n-m)^{-1}h^{5/2} + n'^{-1}h^2\} = O_p(n^{-1}h^2). \tag{B.8}$$

Next, let $K_{ij} \overset{\triangle}{=} K_h(X_{ij}^\top\theta^0)\gamma_0^{-1}(X_i^\top\theta^0)n_c^{-1} = O_p\{(n_ch)^{-1}\}$, we have

$$\begin{aligned}{\sum_{i,s}}'\tilde{\epsilon}_i^{\setminus s}U_{\alpha,i}^\top\sum_{j\notin s}U_{\alpha,j}\epsilon_j &= {\sum_{i,s}}'U_{\alpha,i}^\top\sum_{j\notin s}K_{ij}U_{\alpha,j}\epsilon_j^2 + {\sum_{i,s}}'U_{\alpha,i}^\top\sum_{j_1\notin s,j_2\notin s}^{j_1\neq j_2}K_{ij_1}\epsilon_{j_1}\epsilon_{j_2}U_{\alpha,j_2}\\
&= \binom{n'-2}{m-1}\sum_{(i,j)\in A}^{i\neq j}K_{ij}\epsilon_j^2U_{\alpha,i}^\top U_{\alpha,j} + \binom{n'-1}{m-1}\sum_{i\in A}^{j\notin A}K_{ij}\epsilon_j^2U_{\alpha,i}^\top U_{\alpha,j} + \binom{n'-3}{m-1}\sum_{i\neq j\neq l}^{(i,j,l)\subset A}U_{\alpha,i}^\top U_{\alpha,j}K_{il}\epsilon_j\epsilon_l\\
&+ \binom{n'-2}{m-1}\sum_{i\neq j\neq l}^{(i,j)\subset A,l\notin A}U_{\alpha,i}^\top U_{\alpha,j}K_{il}\epsilon_j\epsilon_l + \binom{n'-1}{m-1}\sum_{j\neq l}^{i\subset A,(j,l)\notin A}U_{\alpha,i}^\top U_{\alpha,j}K_{il}\epsilon_j\epsilon_l\end{aligned}$$

The rate of each term in the above equation can be decided by quantifying corresponding second moments like that in (B.7). Therefore, as $n_ch^2 \to \infty$, we have

$$\frac{1}{2mn_c}\binom{n'}{m}^{-1}{\sum_{i,s}}'\tilde{\epsilon}_i^{\setminus s}U_{\alpha,i}^\top\sum_{j\notin s}U_{\alpha,j}\epsilon_j = O_p(\frac{1}{n^{3/2}h}) = o_p(\frac{1}{n}).$$

which together with (B.8) establishes (B.6). Similarly, we can prove that

$$\frac{2}{m}\binom{n'}{m}^{-1}{\sum_{i,s}}'\{Y_i - \hat{g}_\alpha^{\setminus s}(X_i|\theta_\alpha^0)\}g'(X_{i,\alpha}^\top\theta_\alpha^0)v_\alpha^\top(X_i|\theta^0)\delta_n^{\setminus s} = o_p(\frac{1}{n}), \tag{B.9}$$
$$\frac{2}{m}\binom{n'}{m}^{-1}{\sum_{i,s}}'\{Y_i - \hat{g}_\alpha^{\setminus s}(X_i|\theta_\alpha^0)\}R(X_{i,\alpha}|\theta_\alpha^0, \hat{\theta}_\alpha^{\setminus s}) = o_p(\frac{1}{n}). \tag{B.10}$$

Combining (B.5), (B.6), (B.9) and (B.10), we have

$$\begin{aligned}CV_\alpha(m) &= RSS(m) + m^{-1}\binom{n'}{m}^{-1}\Big\{\frac{1}{n_c^2}{\sum_{i,s}}'U_{\alpha,i}^\top(\sum_{j\notin s}U_{\alpha,j}\epsilon_j)(\sum_{j\notin s}U_{\alpha,j}^\top\epsilon_j)U_{\alpha,i}\\
&\quad + \frac{2}{n_c}{\sum_{i,s}}'\epsilon_iU_{\alpha,i}^\top\sum_{j\notin s}U_{\alpha,j}\epsilon_j\Big\} + o_p(\frac{1}{n}).\end{aligned}$$

Let $T_1 = 2n_c^{-1}\sum_{i,s}{}' \epsilon_i U_{\alpha,i}^\top \sum_{j\notin s} U_{\alpha,j}\epsilon_j$. Then

$$T_1 = \frac{2}{n_c}\binom{n'-2}{m-1}\Big(\sum_i{}'\epsilon_i U_{\alpha,i}^\top \sum_i{}'\epsilon_i U_{\alpha,i} - \sum_i{}'\epsilon_i^2 U_{\alpha,i}^\top U_{\alpha,i}\Big) + \frac{2}{n_c}\binom{n'-1}{m-1}\sum_i{}'\epsilon_i U_{\alpha,i}^\top \sum_{j\notin A} U_{\alpha,j}\epsilon_j.$$

Let $\mathbf{e} = (\epsilon_1, \cdots, \epsilon_n)$, $\mathbf{e}_{s^c} = (\epsilon_j)_{j\notin s}$, $\mathbf{U}_\alpha = (U_{\alpha,1}, \cdots, U_{\alpha,n})^\top$ and $\mathbf{U}_{\alpha,s} = (U_{\alpha,j_1}, \cdots, U_{\alpha,j_m})^\top$, where $j_i \in s$; $\mathbf{U}_{\alpha,s^c} = (U_{\alpha,j_1}, \cdots, U_{\alpha,j_{n_c}})^\top$, where $j_i \notin s$. By (A2) and (A5), we have

$$\begin{aligned}
T_2 &\triangleq \frac{1}{n_c^2}\sum_{i,s}{}' U_{\alpha,i}^\top\Big(\sum_{j\notin s} U_{\alpha,j}\epsilon_j\Big)\Big(\sum_{j\notin s} U_{\alpha,j}^\top\epsilon_j\Big)U_{\alpha,i}\\
&= \frac{1}{n_c^2}\sum_s{}'\Big(\sum_{j\notin s} U_{\alpha,j}^\top\epsilon_j\Big)\sum_{i\in s} U_{\alpha,i}U_{\alpha,i}^\top\Big(\sum_{j\notin s} U_{\alpha,j}\epsilon_j\Big)\\
&= \frac{m}{n_c^2}\sum_s{}'\Big(\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}\Big)(I_{d_\alpha} - \theta_\alpha^0\theta_\alpha^{0\top})\Big(\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}\Big)^\top\\
&= \frac{m}{n_c^2}\sum_s{}'\Big(\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}\Big)\Big(\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}\Big)^\top,
\end{aligned}$$

where the last equation holds as $U_{\alpha,j}^\top\theta_\alpha^0 = 0$ for all j. Simple combinatoric calculation leads to

$$\begin{aligned}
\sum_s{}'\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}\mathbf{U}_{\alpha,s^c}^\top \mathbf{e}_{s^c} &= \binom{n'}{m}\sum_{j\notin A} U_{\alpha,j}^\top\epsilon_j \sum_{j\notin A} U_{\alpha,j}\epsilon_j + 2\binom{n'-1}{m}\sum_{j\notin A} U_{\alpha,j}^\top\epsilon_j \sum_i{}' U_{\alpha,i}\epsilon_i\\
&+ \binom{n'-2}{m}\sum_i{}' U_{\alpha,i}^\top\epsilon_i \sum_i{}' U_{\alpha,i}\epsilon_i + \binom{n'-2}{m-1}\sum_i{}' U_{\alpha,i}^\top U_{\alpha,i}\epsilon_i^2 . \text{(B.11)}
\end{aligned}$$

The coefficient in (B.11) is decided by the following facts

- $\sum_{j\notin A} U_{\alpha,j}\epsilon_j$ is contained in any $\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}$ since $s \subset A$ and there are $\binom{n'}{m}$ such s.
- For any $i \in A$, $\sum_{j\notin A} U_{\alpha,j}^\top\epsilon_j U_{\alpha,i}\epsilon_i$ appears in $\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}$ iff $i \notin s$ and there are $\binom{n'-1}{m}$ such s.
- For any $i_1 \in A$, $i_2 \in A$, the number of s with $\mathbf{e}_{s^c}^\top \mathbf{U}_{\alpha,s^c}$ including term $U_{\alpha,i_1}^\top \epsilon_{i_1} U_{\alpha,i_2}\epsilon_{i_2}$ is $\binom{n'-2}{m}$ if $i_1 \neq i_2$ and $\binom{n'-2}{m}$ if $i_1 = i_2$.

Regarding the terms in (B.11), by weak law of large numbers, we have

$$(\sum_{j\notin A} U_{\alpha,j}^\top \epsilon_j)(\sum_i{}' U_{\alpha,i}\epsilon_i) = O_p\{(n-n')^{1/2}n'^{1/2}\},\ (\sum_{j\notin A} U_{\alpha,j}^\top \epsilon_j)(\sum_{j\notin A} U_{\alpha,j}\epsilon_j) = O_p(n-n'),$$
$$(\sum_i{}' U_{\alpha,i}^\top \epsilon_i)(\sum_i{}' U_{\alpha,i}\epsilon_i) = O_p(n'),\ \sum_i{}' U_{\alpha,i}^\top U_{\alpha,i}\epsilon_i^2 = O_p(n'),$$

The terms involving $\sum\limits_{j\notin A} U_{\alpha,j}\epsilon_j$ are thus negligible compared with the others. Thus

$$\begin{aligned} T_1+T_2 =& \binom{n'-2}{m-1}\frac{1}{n_c^2}\Big\{(2n+n'-3m-1)(\sum_i{}' U_{\alpha,i}^\top\epsilon_i)(\sum_i{}' U_{\alpha,i}\epsilon_i) \\ &+(3m-2n)\sum_i{}' U_{\alpha,i}^\top U_{\alpha,i}\epsilon_i^2\Big\}\{1+o_p(1)\} \\ =& \binom{n'-2}{m-1}\frac{n'}{n_c^2}\Big\{(2n+n'-3m-1)(\frac{1}{\sqrt{n'}}\sum_i{}' U_{\alpha,i}^\top\epsilon_i)(\frac{1}{\sqrt{n'}}\sum_i{}' U_{\alpha,i}\epsilon_i) \\ &+(3m-2n)\frac{1}{n'}\sum_i{}' U_{\alpha,i}^\top U_{\alpha,i}\epsilon_i^2\Big\}\{1+o_p(1)\}. \end{aligned}$$

Note that $n_c = n-m$, both m/n' and m/n tend to c, and

$$\begin{aligned} \frac{n'\binom{n'-2}{m-1}}{m\binom{n'}{m}} &= \frac{n'(n'-2)!m!(n'-m)!}{m(m-1)!(n'-m-1)!n'!} = \frac{(n'-m)}{(n'-1)} \to 1-c, \\ \frac{2n+n'-3m-1}{n_c^2} &= \frac{2n+n'-3m-1}{(n-m)^2} \sim \frac{3}{n-m} \sim \frac{3}{n(1-c)}, \\ \frac{(3m-2n)}{n_c^2} &= \frac{(3m-2n)}{(n-m)^2} \sim \frac{3c-2}{n(1-c)^2}. \end{aligned}$$

By the law of large numbers and (A5),

$$n'^{-1}\sum_i{}' \epsilon_i^2 U_{\alpha,i}^\top U_{\alpha,i} \to \sigma^2 E\{tr(I_{d_\alpha} - \theta^0\theta^{0^\top})\} = \sigma^2(d_\alpha - 1) \text{ in probability.}$$

By the central limit theorem, we have

$$\begin{aligned} &n'^{-1}(\sum_i{}' U_{\alpha,i}^\top\epsilon_i)(\sum_i{}' U_{\alpha,i}\epsilon_i) \xrightarrow{d} \sigma^2\chi^2(d_\alpha-1), \\ &n\Big\{CV_\alpha(m) - RSS(m)\Big\} \to \sigma^2\Big\{3\chi^2(d_\alpha-1) + \frac{(3c-2)(d_\alpha-1)}{(1-c)}\Big\}, \text{ in distribution.} \end{aligned}$$

Recall that $\delta_d = d_\alpha - d_{\alpha_0}$. Since for any $\alpha \supset \alpha_0$, $U_{\alpha_0,i}$ is a subvector of $U_{\alpha,i}$,

$$\Pr\{CV_\alpha(m) - CV_{\alpha_0}(m) > 0\} \to \Pr\{\chi^2(\delta_d) > \frac{(2-3c)\delta_d}{3(1-c)}\},$$

Proof of the consistency for CV(M) in nonparametric models. By simple combinatoric calculations, for nonparametric regression model $E(Y|X) = G(X)$, where $X = (\mathbf{x}_1, \cdots, \mathbf{x}_p)^\top$, we have

$$CV(m) = \sum_{\#s=m}\sum_{i\in s}\{Y_i - \tilde{g}^{\setminus s}(X_i)\}^2 = \sum_{\#\tilde{s}=n-m+1}\sum_{i\in\tilde{s}}\{Y_i - \tilde{g}^{\setminus i}(X_i)\}^2, \tag{B.12}$$

where $\tilde{g}^{\setminus i}(X)$ is the estimate of $g(X)$ from observations indexed by $\tilde{s} \setminus \{i\}$, and $\tilde{s} \subset \{1, \cdots, n\}$ with $\#\tilde{s} = n - m + 1$. Note that the second summation on the right hand side is actually $CV(1)$. If $\alpha \supset \alpha_0$, by Lemma 1 of [89], we have

$$\sum_{i\in\tilde{s}}\{Y_i - \tilde{g}^{\setminus i}(X_{i,\alpha})\}^2 > \sum_{i\in\tilde{s}}\{Y_i - \tilde{g}^{\setminus i}(X_{i,\alpha_0})\}^2 \text{ in probability.} \tag{B.13}$$

Therefore, it follows from (B.12) and (B.13) that $CV_\alpha(m) > CV_{\alpha_0}(m)$ in probability. By (B.12) again and Lemma 1 of [89], if $\alpha \subset \alpha_0$, we also have $CV_\alpha(m) > CV_{\alpha_0}(m)$ in probability. In other words, CV(m) method is consistent.

Proof of Theorem 2.3 We first give the form of CV_α for $\alpha \supseteq \alpha_0$ and $\alpha \cup p = \alpha_0$.

1. $\alpha \supseteq \alpha_0$. In the proof of Theorem 2.2, when $m = 1$ we know that

   $$CV_\alpha = \frac{1}{n'}{\sum_i}'\{Y_i - \hat{g}_\alpha^{\setminus i}(X_i|\theta_\alpha^0)\}^2 + O_p(\frac{1}{n}).$$

   Since $n - n' = O(nh)$ and that the local linear estimator is used, by (B.4) we have $Y_j - \hat{g}_\alpha^{\setminus j}(X_j|\theta_\alpha^0) = \epsilon_j + O_p(\tau_n)$ uniformly in $j \notin A$, where the term $O_p(\tau_n)$ is independent of ϵ_j. Therefore,

   $$\sum_{j\notin A}\{Y_j - \hat{g}_\alpha^{\setminus j}(X_j|\theta_\alpha^0)\}^2 = \sum_{j\notin A}\epsilon_j^2 + O_p(\log n).$$

We can write

$$\sum_i{}' \{Y_i - \hat{g}_\alpha^{\setminus i}(X_i|\theta_\alpha^0)\}^2 = \sum_{i=1}^n \{Y_i - \hat{g}_\alpha^{\setminus i}(X_i|\theta_\alpha^0)\}^2 - \sum_{j\notin A} \{Y_j - \hat{g}_\alpha^{\setminus j}(X_j|\theta_\alpha^0)\}^2.$$

Following the steps in the proof of Lemma 1 in [89], we have

$$n^{-1}\sum_{i=1}^n \{Y_i - \hat{g}_\alpha^{\setminus i}(X_i|\theta_\alpha^0)\}^2 = n^{-1}\sum_{i=1}^n \epsilon_i^2 + c_1(nh)^{-1} + c_2h^4 + o_p\{(nh)^{-1}\},$$

where $c_1 = \sigma^2 K_2 E\{\gamma_0^{-1}(X^\top\theta^0)\}$, $c_2 = Eg''^2(X^\top\theta^0)/4$. Recall that $\mathcal{U} = \{X^\top\theta^0 : X \in A\}$. Note that since $\alpha \supset \alpha_0$, $\mathcal{U}$ also equals to $\{X_\alpha^\top\theta_\alpha^0 : X \in A\}$. By (A1), $c_1 = \sigma^2K_2L(\mathcal{U})$ with $L(\mathcal{U})$ being the Lebesgue measure of $\mathcal{U}$. Thus,

$$CV_\alpha = \frac{1}{n'}\sum_i{}' \epsilon_i^2 + \frac{c_1}{n'h} + c_2h^4 + o_p(\frac{1}{n'h}). \tag{B.14}$$

2. $\alpha \cup d = \alpha_0$. Let $\hat{g}_\alpha(X_i|\theta)$ be defined similarly to (2.8) but using all observations. Then by Lemma B.1, $\hat{g}_\alpha(X_i|\hat{\theta}_\alpha^{\setminus i}) - \hat{g}_\alpha^{\setminus i}(X_i|\hat{\theta}_\alpha^{\setminus i}) = O_p(\tau_n)$ uniformly in i. Therefore, by Theorem 6 in [54], for any $\theta \in \Theta_{n,\alpha}$,

$$\max_{X\in A}\left|\hat{g}_\alpha^{\setminus i}(X_\alpha|\theta) - g_\alpha(X_\alpha|\theta)\right| = O_p(\tau_n). \tag{B.15}$$

Step 2 in the algorithm indicates that $\hat{\theta}_\alpha^{\setminus i}$ is the first d_α entry of the MAVE estimator of SIM: $Y = g(X_{\alpha\cup d}^\top\theta) + \epsilon$ using data $\{X_j, Y_j\}_{j\neq i}$. Therefore, by (B.1), $\hat{\theta}_\alpha^{\setminus i} - \theta_\alpha^0 = O_p(n^{-1/2})$ uniformly in i. By (B.15) with θ replaced by θ_α^0 and $\hat{\theta}_\alpha^{\setminus i}$, we have

$$\begin{aligned}\hat{g}_\alpha^{\setminus i}(X_\alpha|\theta_\alpha^0) - \hat{g}_\alpha^{\setminus i}(X_\alpha|\hat{\theta}_\alpha^{\setminus i}) =& \hat{g}_\alpha^{\setminus i}(X_\alpha|\theta_\alpha^0) - g_\alpha(X_\alpha|\theta_\alpha^0) + g_\alpha(X_\alpha|\theta_\alpha^0) - g_\alpha(X_\alpha|\hat{\theta}_\alpha^{\setminus i})\\ &+ g_\alpha(X_\alpha|\hat{\theta}_\alpha^{\setminus i}) - \hat{g}_\alpha^{\setminus i}(X_\alpha|\hat{\theta}_\alpha^{\setminus i})\\ =& O_p(\tau_n), \text{ uniformly in } i,\end{aligned}$$

Hence $Y_i - \hat{g}_\alpha^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i}) = Y_i - \hat{g}_\alpha^{\backslash i}(X_i|\theta_\alpha^0) + O_p(\tau_n)$ uniformly in i with term $O_p(\tau_n)$ independent of $\{X_i, Y_i\}$. Therefore, according to Lemma 1 in [89],

$$CV_\alpha = n'^{-1}\sum_i \{Y_i - \hat{g}_\alpha^{\backslash i}(X_i|\theta_\alpha^0)\}^2 + o_p(1) \xrightarrow{p} \sigma_\alpha^2(\theta_\alpha^0). \tag{B.16}$$

The form of $CV_{\alpha,d}$ with $\mathbf{x}_d$ discrete is different to that with $\mathbf{x}_d$ continuous.

1. For discrete x_d with M values $v_1, \cdots, v_M$, we classify $\{(X_i, Y_i)\}_{i=1}^{n'}$ into M groups based on the value of $x_d : i \in G_k \Leftrightarrow \mathbf{x}_{id} = v_k$. Let n_k be the number of elements in G_k and $n_k = O(n'),\ k = 1, \cdots, M$. If $i \in G_k$, by (2.11), $\hat{g}_{\alpha,d}^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i})$ equals to $\hat{g}_\alpha^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i})$, which is defined in (2.8) with θ replaced by $\hat{\theta}_\alpha^{\backslash i}$ and subindex $\{j \notin s\}$ by $\{j \in G_k, j \neq i\}$. Thus $CV_{\alpha,d} = n'^{-1}\sum_{k=1}^M n_k CV_\alpha^k$, where $CV_\alpha^k{:=}{n_k}^{-1}\sum_{i\in G_k}\{Y_i - \hat{g}_\alpha^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i})\}^2$ is the $CV_\alpha(1)$ in (2.9) using data $\{(X_{i\alpha}, Y_i) : i \in G_k\}$. Since $\alpha \cup d \supseteq \alpha_0$, $E(Y|X)$ depends only on X_α within each G_k. Therefore, similarly to (B.14), by (A7) we then have

$$CV_\alpha^k = \frac{1}{n_k}\sum_{i\in G_k}\epsilon_i^2 + c_4 h_1^4 + \frac{\sigma^2 K_2}{n_k h_1} L(\mathcal{U}_\alpha^k) + o_p(\frac{1}{n_k h_1}),\ k = 1, \cdots, M,$$

where $c_4 = E\{g''^2(X_\alpha^\top \theta_\alpha^0)|x_d = v_k\}/4$, and $\mathcal{U}_\alpha^k$ is the support of $X_\alpha^\top\theta_\alpha^0$ given that $x_d = v_k$. Therefore,

$$CV_{\alpha,d} = \frac{1}{n'}{\sum_i}'\epsilon_i^2 + \frac{\sigma^2 K_2}{n' h_1}\sum_{k=1}^M L(\mathcal{U}_\alpha^k) + c_2 h_1^4 + o_p(\frac{1}{n' h_1}). \tag{B.17}$$

Note that if x_d is redundant, i.e. $\theta_d^0 = 0$, then $\mathcal{U}_\alpha^k$ is also the support of $X^\top\theta^0$ given that $\mathbf{x}_d = v_k$. By the discussion about the identification of single-index models with discrete covariates ([44]), we have $\sum_{k=1}^M L(\mathcal{U}_\alpha^k) > L(\mathcal{U})$.

2. $\mathbf{x}_d$ continuous: Note that if $\alpha \cup d \supseteq \alpha_0$, then $g_{\alpha,d}(u, v|\theta_\alpha^0) = g(u^\top \theta_\alpha^0 + \theta_d^0 v)$. Similarly to (B.4), we have

$$\begin{aligned}\hat{g}_{\alpha,d}(u, v|\theta) = & g(u, v|\theta_\alpha^0) + g'(u^\top \theta_\alpha^0 + \theta_d^0 v)(\theta_\alpha^0 - \theta)^\top v_\alpha(u|\theta^0) + \frac{1}{2} g''(u^\top \theta_\alpha^0 + \theta_d^0 v) h_1^2 \\ & + \frac{1}{n} f^{-1}_{X_\alpha^\top \theta, x_d}(u^\top \theta_\alpha^0, v) \sum_{i=1}^n K_{h_1}((X_{i,\alpha} - u)^\top \theta_\alpha^0) H_{h_1}(\mathbf{x}_{i,d} - v)\epsilon_i + o_p(n^{-1/2}).\end{aligned}$$

Following the proof of Theorem 2.2, we have

$$\hat{g}_{\alpha,d}^{\backslash i}(u, v|\theta_\alpha^0) - \hat{g}_{\alpha,d}^{\backslash i}(u, v|\hat{\theta}_\alpha^{\backslash i}) = g'(u^\top \theta_\alpha^0 + \theta_d^0 v)(\hat{\theta}_\alpha^{\backslash i} - \theta_\alpha^0)^\top v_\alpha(u|\theta^0) + O_p(h_1^2 \log^{1/2} n),$$

uniformly in i and $u \in \mathcal{D}_\alpha,\ v \in A_d$. Therefore,

$$\begin{aligned}Y_i - \hat{g}_{\alpha,d}^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i}) = & Y_i - \hat{g}_{\alpha,d}^{\backslash i}(X_i|\theta_\alpha^0) + \hat{g}_{\alpha,p}^{\backslash i}(X_i|\theta_\alpha^0) - \hat{g}_{\alpha,d}^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i}) \\ = & Y_i - \hat{g}_{\alpha,d}^{\backslash i}(X_i|\theta_\alpha^0) + \frac{1}{n} U_{\alpha,i}^\top \sum_{j \neq i} U_{\alpha,j} + \text{Higher order terms} \\ \overset{\triangle}{=} & T_{1i} + T_{2i} + \text{higher order terms.}\end{aligned}$$

Similar arguments as in the proof Theorem 2.2 can be engaged to deal with $\sum_i{}' T_{1i} T_{2i}$ and $\sum_i{}' T_{2i}^2$. Finally, again through arguments similar to that in [89], we have

$$\begin{aligned}CV_{\alpha,d} = & \frac{1}{n'} \sum_i{}' \{Y_i - \hat{g}_{\alpha,d}^{\backslash i}(X_i|\theta_\alpha^0)\}^2 + o_p(\frac{1}{n' h_1^2}) & \text{(B.18)} \\ = & \frac{1}{n'} \sum_i{}' \epsilon_i^2 + \frac{c_5}{n' h_1^2} + E\Big\{g''^2(X^\top \theta^0)\Big\} h_1^4 + o_p(\frac{1}{n' h_1^2}), & \text{(B.19)}\end{aligned}$$

where $c_5 = \sigma^2 K_2^2 L(X_\alpha^\top \theta_\alpha^0) L(A_d) > 0$.

Note that the bandwidth $h \propto n^{-1/5}$ in CV_α, and in $CV_{\alpha,d}$, $h_1 = h$ for x_d discrete and $h_1 \propto n^{-1/6}$ for x_d continuous. Comparing $CV_{\alpha,d}$ in (B.17) and (B.19) with CV_α in (B.14) and (B.16), we complete the proof.

Appendix C

Conditions and Proofs for Chapter 3

we first introduce some notation. Let

$$\mu_i = \int t^i K(t)dt,\ \mu_i^* = \int t^i K^2(t)dt,\ \Gamma = \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_2 \end{pmatrix},\ \Gamma^* = \begin{pmatrix} \mu_0^* & 0 \\ 0 & \mu_2^* \end{pmatrix}$$

$$\sigma^2(X_i, V_i) = Var(\xi_i),\ \Omega(v) = [\omega_{i,j}(v)] = E(X_i X_i^\top | V_i = v),$$

$$\Omega^*(v) = [\omega_{i,j}^*(v)] = E\Big\{X_i X_i^\top \sigma^2(X_i, V_i) | V_i = v\Big\},\ G = I_{p+1} \otimes \mathrm{diag}(1, h).$$

$$C(v) = f(v)\Omega(v) \otimes \Gamma,\ \xi = (\xi_1, \cdots, \xi_n)^\top,\ Y = (Y_1^2, \cdots, Y_n^2)^\top.$$

Let $\delta_n = (nh/\ln n)^{-1/2}$ and Ω_j is the jth column of Ω. Let $\mathcal{F}_i^k$ be the $\sigma-$algebra generated by $\{V_j, Y_j\}_{j=i}^k$. The process $\{V_j, Y_j\}$ is $\alpha-$mixing if the mixing coefficient

$$\alpha(k) := \sup_{\substack{A \in \mathcal{F}_{-\infty}^0 \\ B \in \mathcal{F}_k^\infty}} |P(AB) - P(A)P(B)| \to 0,\ \text{as } k \to \infty.$$

Among various mixing conditions used in time series literature, $\alpha-$mixing is reasonably weak, and is known to be fulfilled for many stochastic processed including many times series models. [4] provided illuminating discussions on the role of $\alpha-$mixing (including geometric ergodicity) for model identification in nonlinear time series analysis. Further, [55, 56] showed that under some mild conditions, both ARCH process and NAARX (additive autoregressive process with exogenous variables) are stationary and $\alpha-$mixing.

(A1) The function $K(.)$ is a symmetric and bounded density with a bounded support.

(A2) The density function $f(v)$ of V is bounded from 0 on its compact support $\mathcal{D}$ with bounded first-order derivative.

(A3) $\Omega(v)$ is nonsingular for all $v \in \mathcal{D}$. $\Omega(v)$ and $\Omega^{-1}(v)$ have bounded-first order derivatives.

(A4) $EY^{2\delta^*} < \infty$, for some $\delta^* > 2$.

(A5) The coefficient functions $a_k(.)$, $k = 0, \cdots, p$ all have second-order derivatives in $\mathcal{D}$ and are *Lipschitz* continuous $|a_k^{''}(v_1) - a_k^{''}(v_2)| \leq c|v_1 - v_2|$, for some $c > 0$.

(A6) Let $\mathsf{Y}_l := (Y_{lp}^2, \cdots, Y_{l1}^2)$. The conditional density $f(V_l, \mathsf{Y}_l | Y_l)$ of (V_l, Y_l) given Y_l exists and bounded; the conditional density $f(\mathsf{Y}_1, v_1, \mathsf{Y}_l, v_l | Y_1, Y_l)$ of $(\mathsf{Y}_1, v_1, \mathsf{Y}_l, v_l)$ given (Y_1, Y_l) exists and bounded for all $l \geq 1$.

(A7) The processes $\{V_t, Y_t\}$ are stationary and $\alpha-$mixing with

$$\sum k^c \{\alpha(k)\}^{1-2/\delta} < \infty, \quad \sum_{n=1}^{\infty} \psi(n) < \infty, \tag{C.1}$$

for some $2 < \delta \leq \delta^*$ and $c > 1 - 2/\delta$, where

$$\psi(n) := \frac{nL(n)}{r(n)} \left(\frac{nT_n^2}{h \ln n}\right)^{1/4} \alpha\{r(n)\}, \quad r(n) := \frac{(nh/\ln n)^{1/2}}{T_n}$$
$$L(n) := \left(\frac{nT_n^2}{h^3 \ln n}\right)^{1/2}, \quad T_n := \{n \ln n (\ln\ln n)^{1+\mu}\}^{1/\delta} \text{ for some } 0 < \mu < 1.$$

(A8)The bandwidth $h \to 0$ with $n^{2/\delta^*-1} h^{-1} \ln n^{1+2/\delta^*} (\ln\ln n)^{2(1+\delta)/\delta^*} \longrightarrow 0$.

(A1)-(A6) are regular assumptions for regression models in time series analysis. Conditions (A7)(A8) on $\alpha(k)$ and h are required to ensure the strong uniform convergence rate of local linear estimators for time series ([54]); see [54] for an explicit rate of decay for $\alpha(k)$ of the form $\alpha(k) = O(1/k^c)$ for some $c > 0$. For local linear estimators in varying coefficient regression models for nonlinear time series, the asymptotic normality (3.16) was proved by [11] to hold under conditions weaker than in [54].

Proof of Theorem 3.1 By Taylor's expansion around v, we have

$$Y = \tilde{X}_0(a_0(v), a_0'(v), \cdots, a_p(v), a_p'(v))^\top + \frac{1}{2}\sum_{j=0}^{p} \begin{pmatrix} a_j^{''}(\varepsilon_{1j})V_{1v}^2 Y_{1j}^2 \\ \vdots \\ a_j^{''}(\varepsilon_{nj})V_{nv}^2 Y_{nj}^2 \end{pmatrix} + \xi$$

where $\tilde{X}_0$, and W_0 are similarly defined as $\tilde{X}_i$, W_i in (3.17) with v^i replaced by v. and ε_{ij} lies between V_i and v, for $i = 1, \cdots, n$, $j = 0, \cdots, p$. Therefore,

$$\hat{a}_j(v) = a_j(v) + \frac{1}{2}\sum_{j=0}^{p} e_{2j+1,2p+2}^\top (\tilde{X}_0^\top W \tilde{X}_0)^{-1} \tilde{X}_0^\top W \begin{pmatrix} a_j^{''}(\varepsilon_{1j})V_{1v}^2 Y_{1j}^2 \\ \vdots \\ a_j^{''}(\varepsilon_{nj})V_{nv}^2 Y_{nj}^2 \end{pmatrix}$$
$$+ e_{2j+1,2p+2}^\top (\tilde{X}_0^\top W \tilde{X}_0)^{-1} \tilde{X}_0^\top W \xi, \quad j = 0, \cdots, p. \tag{C.2}$$

Therefore, by Lemma C.1, Lemma C.2 and (C.5), we have from (C.2)

$$\hat{a}_j(v) - a_j(v) = \frac{h^2}{2} e_{2j+1,2p+2}^\top G^{-1} C^{-1}(v) \Big\{ \sum_{j=0}^{p} a_j^{''}(v) \omega_{j+1} \otimes \begin{pmatrix} \mu_2 \\ 0 \end{pmatrix} \Big\} + O_p(\delta_n)$$

uniformly for all $v \in \mathcal{D}$. By the property of Kronecker product, $C^{-1}(v)$ and $a_j^{''}(v)\omega_{j+1} \otimes (1,0)^\top$ admit the following two forms respectively:

$$\begin{pmatrix} \star & 0 & \cdots & \star & 0 \\ 0 & \star & \cdots & 0 & \star \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \star & 0 & \cdots & \star & 0 \\ 0 & \star & \cdots & 0 & \star \end{pmatrix} \qquad \begin{pmatrix} \star \\ 0 \\ \vdots \\ \star \\ 0 \end{pmatrix} \tag{C.3}$$

which completes the proof.

Lemma C.1 facilitates the approximation of the random matrices on the right hand side of (C.2) by the corresponding deterministic ones.

Lemma C.1 *Suppose (A1),(A4) and (A6)-(A8) hold. Then*

$$\sup_{v\in\mathcal{D}} \left| \frac{1}{n} G^{-1}\tilde{X}_0^\top W_0 \tilde{X}_0 G^{-1} - C(v) \right| = O_p(\delta_n + h). \tag{C.4}$$

$$\sup_{v\in\mathcal{D}} \left| \frac{1}{n} G^{-1}\tilde{X}_0^\top W \begin{pmatrix} a_j^{''}(\varepsilon_{1j})V_{1v}^2 Y_{1j}^2 \\ \vdots \\ a_j^{''}(\varepsilon_{nj})V_{nv}^2 Y_{nj}^2 \end{pmatrix} - f(v)h^2 a_j^{''}(v)\Omega_{j+1} \otimes \begin{pmatrix} \mu_2 \\ 0 \end{pmatrix} \right| = O_p(\tau_n),\ 0 \le j \le p.$$

Proof. We only prove the first equation for illustration. Note that for $1 \le l, t \le 2,\ 0 \le k, s \le p$, the $(2k+l, 2s+t)$ position element of $n^{-1}G^{-1}\tilde{X}_0^\top W \tilde{X}_0 G$ is given by

$$B_{kl,st} := \frac{1}{n} h^{2-t-l} \sum_{j=1}^n Y_{jk}^2 Y_{js}^2 V_{jv}^{t+l-2} K_{h,j}(v).$$

where Y_{jk} is defined in right above (3.17). By Lemma 3 in [54],

$$h^{2-t-l} \left| B_{kl,st} - E\Big\{ Y_{jk}^2 Y_{js}^2 V_{jv}^{t+l-2} K_{h,j}(v) \Big\} \right| = O_p(\delta_n),$$

uniformly for all $v \in \mathcal{D}$. Note that $E\Big\{ Y_{jk}^2 Y_{js}^2 V_{jv}^{t+l-2} h^{2-t-l} K_{h,j}(v) \Big\}$ does not depend on j from (A7). By (A2) and (A4), we have

$$E\Big\{ Y_{jk}^2 Y_{js}^2 (\frac{V_{jv}}{h})^{t+l-2} K_{h,j}(v) \Big\} = [\Omega_{(k+1,s+1)} f](v)\mu_{t+l-2} + h\mu_{t+l-1}[\Omega_{(k+1,s+1)} f]'(v) + O(h^2),$$

uniformly for $v \in \mathcal{D}$. The proof is thus complete since the $(2k+l, 2s+t)$ position element of $C(v)$ is exactly $\Omega_{(k+1,s+1)}(v)f(v)\mu_{t+l-2}$.

By (C.4) and the fact that $(A+hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$, we have

$$nG(\tilde{X}_0^\top W\tilde{X}_0)^{-1}G = C^{-1}(v) + O_p(\delta_n + h) \tag{C.5}$$

uniformly for $v \in \mathcal{D}$.

Lemma C.2 *Let $\tau_n = h^2(\delta_n + h)$ and suppose (A8) holds. Then*

$$\sup_{v\in\mathcal{D}} \left|\frac{1}{n}G^{-1}\tilde{X}_0^\top W\xi\right| = O_p(\tau_n) \tag{C.6}$$

Proof By definition, the $2k+l(l \leq 1)$th component of $n^{-1}G^{-1}\tilde{X}_0^\top W\xi$ is given by $n^{-1}\sum_{i=1}^n Y_{ik}^2(V_{iv}/h)^{l-1}K_{h,i}(v)\xi_i$. Since given V_i and V_{i+l}, ϵ_i and ϵ_{i+l} are independent, the mixing coefficient of the process $Y_{ik}^2\xi_i$ still satisfies (C.1). (C.6) thus follows from Lemma 3 in [54] and the fact that $E(\xi_i|V_i) = 0$.

Proof of Theorem 3.2 According to [49], (3.21) can be represented as the following fundamental problem, which can then be quickly soved by the Lemke([47]) or Dantzig-Cottle([18, 15]) algorithms

$$v_n = W_n\lambda_n + q_n \quad \text{subject to} \quad v_n^\top\lambda_n = 0,\ v_n \geq \mathbf{0},\ \lambda \geq \mathbf{0}, \tag{C.7}$$

where $W_n = AQ^{-1}A^\top$, $q_n = A\hat{\alpha}$, $\hat{\alpha} = Q^{-1}C$. Let v_n, λ_n be the nonnegative complementary solution to $(C.7)$. The solution to (3.21) is thus given by

$$\tilde{\alpha} = \hat{\alpha} + Q^{-1}A\lambda_n. \tag{C.8}$$

Case 1: $A\alpha > \mathbf{0}$. Based on (3.16) and Theorem 3.1, both (3.22) and (3.23) are true if we can prove that $P(\tilde{\alpha} - \hat{\alpha} \neq 0) \to 0,\ n \to \infty$. To do this, note that the Dantzig-Cottle system (C.7) implies that if $q_n > \mathbf{0}$, then $\lambda_n = \mathbf{0}$ and consequently $\tilde{\alpha} = \hat{\alpha}$ by (C.8). Write $q_n = A\hat{\alpha} = A\alpha + A(\hat{\alpha} - \alpha)$. It's easy to see that $q_n > \mathbf{0}$ in probability, since $A\alpha > \mathbf{0}$ and $A(\hat{\alpha} - \alpha) = o_p(1)$ by (3.16). The proof is thus complete.

Case 2: $A_1\alpha > \mathbf{0}$, **and** $A_2\alpha = \mathbf{0}$.

According to [49], the solution to $\min_\alpha(\alpha^\top Q\alpha - 2C^\top\alpha)$ subject to $A_2\alpha = \mathbf{0}$ is

$$\alpha^* = \hat{\alpha} - Q^{-1}A_2^\top \left(A_2 Q^{-1} A_2^\top\right)^{-1} A_2\hat{\alpha} = \hat{\alpha} - Q^{-1}A_2^\top \left(A_2 Q^{-1} A_2^\top\right)^{-1} A_2(\hat{\alpha} - \alpha),$$

Let $\Sigma = diag\{C^{-1}(v^1), \cdots, C^{-1}(v^m)\}$, $\Xi = \Sigma A_2^\top(A_2\Sigma A_2^\top)^{-1}A_2$. Then by Lemma C.1

$$Q^{-1} = \frac{1}{n}\mathrm{diag}\{\mathbf{1}_{m\times 1} \otimes G^{-1}\}\ \Sigma\ \mathrm{diag}\{\mathbf{1}_{m\times 1} \otimes G^{-1}\}\{1 + O_p(\delta_n + h)\}.$$

As $C(v^i)$ admits the form in (C.3) and all the even-numbered columns of A_2 are zero vectors, $Q^{-1}A_2^\top(A_2Q^{-1}A_2^\top)^{-1}A_2 = \Xi\{1 + O_p(\delta_n + h)\}$. Therefore,

$$\alpha^* - \hat{\alpha} = -\Xi(\hat{\alpha} - \alpha)\{1 + O_p(\delta_n + h)\}.$$

By Lemma C.3, we can see that (3.22) follows from Theorem 3.1 and the fact that

$$\tilde{\alpha} - \hat{\alpha} = -\Xi(\hat{\alpha} - \alpha)\{1 + O_p(\delta_n + h)\},$$

$$\tilde{\alpha} - \alpha = (\mathbf{I} - \Xi)(\hat{\alpha} - \alpha)\{1 + O_p(\delta_n + h)\}.$$

Let $\mathbf{a} = \{\mathbf{a}(v^1)^\top, \cdots, \mathbf{a}(v^m)^\top\}^\top$, $\hat{\mathbf{a}} = \{\hat{\mathbf{a}}(v^1)^\top, \cdots, \hat{\mathbf{a}}(v^m)^\top\}^\top$, $\mathbf{a}'' = \{\mathbf{a}''(v^1)^\top, \cdots, \mathbf{a}''(v^m)^\top\}^\top$. Using methods in [11] to prove (3.16), we can prove that

$$(nh)^{1/2}\Big\{\hat{\mathbf{a}} - \mathbf{a} - \frac{\mu_2}{2}h^2\mathbf{a}''\Big\} \to N(0, \Theta),\ \Theta := \mathrm{diag}\{\Theta_1(v^1), \cdots, \Theta_1(v^m)\}. \tag{C.9}$$

Let J_i be a $(p+1)\times m(2p+2)$ matrix with $J_i(k+1, (2p+2)(i-1)+2k+1) = 1,\ k = 0, \cdots, p$, and other entries zero, and Ξ_i a $(p+1) \times m(p+1)$ matrix with

$$\Xi_i(k, l) = \Xi((2p+2)(i-1) + 2k - 1, (2p+2)(i-1) + 2l - 1).$$

Then it is easy to check that

$$\hat{\mathbf{a}}(v^i) - \mathbf{a}(v^i) = J_i(\hat{\alpha} - \alpha),\ \tilde{\mathbf{a}}(v^i) - \mathbf{a}(v^i) = J_i(\tilde{\alpha} - \alpha),\ J_i\Xi(\hat{\alpha} - \alpha) = \Xi_i(\hat{\mathbf{a}} - \mathbf{a}),$$

where the last equation holds by the fact that all the even-numbered columns and rows of Ξ are zero vectors. Therefore,

$$\tilde{\mathbf{a}}(v^i) - \mathbf{a}(v^i) = \hat{\mathbf{a}}(v^i) - \mathbf{a}(v^i) - \Xi_i(\hat{\mathbf{a}} - \mathbf{a}) = (H_i - \Xi_i)(\hat{\mathbf{a}} - \mathbf{a})\{1 + O_p(\delta_n + h)\},$$

where H_i is a $(p+1) \times m(p+1)$ matrix with $H_i(k, (i-1)(p+1)+k) = 1,\ k = 1, \cdots, p+1$, and other entries zero. Finally, by (C.9),

$$(nh)^{1/2}\Big\{\tilde{\mathbf{a}}(v^i) - \mathbf{a}(v^i) - \frac{\mu_2}{2}h^2(H_i - \Xi_i)\mathbf{a}^{''}(v^i)\Big\} \to N\Big\{0, (H_i - \Xi_i)\Theta(H_i - \Xi_i)^{\top}\Big\}.$$

Lemma C.3 To prove $P(\tilde{\alpha} - \alpha^* \neq 0) \to 0$, as $n \to \infty$, first note that for any constrained least-square problem

$$\min_b Z = \tfrac{1}{2}(y - Xb)^{\top}(y - Xb) \qquad \text{(C.10)}$$

$$Ab \geq c \quad (\text{ or } Ab - v = c) \qquad \text{(C.11)}$$

(where v is a surplus vector and b is not otherwise restricted) is equivalent to

$$\max_{\lambda} L = c'\lambda + \tfrac{1}{2}(y'y - b'X'Xb)$$

$$A'\lambda + X'y = (X'X)b, \quad \lambda \geq \mathbf{0}, \qquad \text{(C.12)}$$

where λ is a dual vector and b is the solution to (C.10). (C.11) and (C.12) can thus be partitioned as

$$A_1\alpha - v_1 = \mathbf{0}, \quad A_2\alpha - v_2 = \mathbf{0},$$

$$(A_1' \ \ A_2')\begin{pmatrix}\lambda_1\\ \\ \lambda_2\end{pmatrix} + C = Q\alpha$$

or

$$\tilde{\alpha} = Q^{-1}A_1'\lambda_1 + Q^{-1}A_2'\lambda_2 + Q^{-1}C. \tag{C.13}$$

By the Dantzig-Cottle conditions,

$$v_1'\lambda_1 = 0,\ v_2'\lambda_2 = 0,\ v_1 \geq \mathbf{0},\ \lambda_1 \geq \mathbf{0},\ v_2 \geq \mathbf{0},\ \lambda_2 > \mathbf{0}.$$

Through the arguments in [49], we have

$$\begin{aligned}
&\lambda_2 = -W_{22}^{-1}W_{21}\lambda_1 - W_{22}^{-1}A_2\hat{\alpha}, \quad \hat{\alpha} = Q^{-1}C,\\
&v_1 = M^*\lambda_1 + q_n, \quad M^* = W_{11} - W_{12}W_{22}^{-1}W_{21},
\end{aligned}$$

where $W_{ij} = A_iQ^{-1}A_j',\ i,j = 1,2$, and

$$\begin{aligned}
q_n &= -W_{12}W_{22}^{-1}A_2\hat{\alpha} + A_1\hat{\alpha}\\
&= A_1\alpha - W_{12}W_{22}^{-1}A_2\alpha + (A_1 - W_{12}W_{22}^{-1}A_2)(\hat{\alpha} - \alpha)
\end{aligned}$$

By the prior belief $A_1\alpha > 0$ and $A_2\alpha = 0$ and Theorem 3.1, $q_n > \mathbf{0}$ in probability. Since $q_n > 0$ implies $v_1 = q_n > 0$ and $\lambda_1 = 0$, we have

$$\lambda_2 = -(A_2Q^{-1}A_2')^{-1}A_2\hat{\alpha}. \tag{C.14}$$

Substitute λ_2 in (C.13) for (C.14) and we have

$$\tilde{\alpha} = \hat{\alpha} - Q^{-1}A_2'(A_2Q^{-1}A_2')^{-1}A_2\hat{\alpha} = \alpha^*.$$