I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2012/25

EFFICIENT MODEL SELECTION IN SEMIVARYING COEFFICIENT MODELS

NOH, H. and I. VAN KEILEGOM

Efficient Model Selection in Semivarying Coefficient Models

Hohsuk Noh
Université catholique de Louvain

Ingrid Van Keilegom
Université catholique de Louvain

September 1, 2012

Abstract

Varying coefficient models are useful extensions of classical linear models. In practice, some of the varying coefficients may be just constant, while other coefficients are varying. Several methods have been developed to utilize the information that some coefficient functions are constant to improve estimation efficiency. However, in order for such methods to really work, the information about which coefficient functions are constant should be given in advance. In this paper, we propose a computationally efficient method to discriminate in a consistent way the constant coefficient functions from the varying ones. Additionally, we compare the performance of our proposal with that of previous methods developed for the same purpose in terms of model selection accuracy and computing time.

1 Introduction

Varying coefficient models are an important generalization of classical linear models. They form an alternative class of models to ameliorate the so-called 'curse of dimensionality' in multidimensional nonparametric modeling from a theoretical point of view, but they also allow us to explore the dynamic features hidden in the data. Let Y be the response variable and $(U, X_1, \ldots, X_p)$ be its associated covariates. Then the varying coefficient model is defined by the following linear model:

$$Y = \sum_{l=1}^{p} a_l(U) X_l + \epsilon, \tag{1.1}$$

with $E(\epsilon|U, X_1, \ldots, X_p) = 0$ and $\mathrm{Var}(\epsilon|U, X_1, \ldots, X_p) = \sigma^2(U)$. In practice, occasionally, some of the coefficient functions in model (1.1) are constant, while other coefficients are varying with respect to

the index variable U. In this case, we can rewrite the model as

$$Y = \sum_{l=1}^{s} a_l(U)X_l + \sum_{l=s+1}^{p} a_l X_l + \epsilon, \tag{1.2}$$

which is the so-called semi-varying coefficient model. From an estimation point of view, this model cannot be treated as a special case of a varying coefficient model, because treating constant coefficients as varying will result in loss of estimation efficiency. This motivated several authors to develop methods to incorporate the information of constancy into the estimation procedure. To obtain a root-n convergence rate for the constant coefficients as in parametric models, Zhang et al. (2002) and Cheng et al. (2009) considered a two-step estimation procedure that estimates the constant coefficients by taking the average of the initial estimators over a grid of points. Xia et al. (2004) proposed a semi-local least squares method that estimates the varying coefficients locally and the constant coefficients globally. Additionally, Fan and Huang (2005) showed that it is also possible to have a parametric convergence rate for the constant coefficients with the profiled least squares method. However, in order for such methods to really work, the information about which coefficient functions are constant should be given in advance. Since such information is rarely available in practice, it is of interest to develop an efficient and fast method to discriminate constant coefficients from varying ones.

Identifying constant coefficients can be done through hypothesis testing as in Fan et al. (2001), Fan and Huang (2005) and Wang et al. (2009). Nevertheless, the theoretical properties of such identification based on hypothesis testing can be somewhat hard to analyze. Another way to identify the constant coefficients is to consider the problem in a variable selection framework. This type of approach has been considered often for the identification of nonzero coefficients in varying coefficient models as in Wang et al. (2008), Noh and Park (2010) and Antoniadis et al. (2012) to name a few but not as frequent in our context. Xia et al. (2004) proposed a cross-validation (CV) procedure based on local linear estimators to discriminate constant coefficients from varying ones. However, the implementation of their method is computationally very demanding. Indeed, to calculate the CV score for each candidate model, the bandwidth should be chosen via leave-one-out cross-validation using the estimates based on the semi-local least squares method involving large-scale matrix computation. Further, the computation becomes even more challenging when the number of covariates increases. To tackle this problem, Hu and Xia (2012) addressed the same problem as a shrinkage estimation problem using local polynomial smoothing and proposed a Bayesian Information Criterion to select the shrinkage parameter. Although the method is able to identify the constant coefficients without

doing an exhaustive search over the candidate models, its discrimination ability is not as sensible as for the method of Xia et al. (2004), because it is based on local constant estimators. Additionally, the adaptation of the method of Hu and Xia (2012) to local linear estimation is not at all a trivial task. The reason is that when the coefficient function $a_l(\cdot)$ is constant we should estimate $a_l(\cdot)$ as constant but its derivative $a_l'(\cdot)$ as zero. Due to this, we should consider a totally different penalty from what Hu and Xia (2012) considered and hence much theoretical and computational work should be done for the extension. However, the extension is necessary, since local constant estimators tend to suffer from boundary problems, which can substantially degrade the discrimination ability of the procedure. Based on this observation, we are motivated to develop a new method, whose discrimination ability is as sensible as for the procedure of Xia et al. (2004), which is based on local linear estimators, and which is as fast as the one of Hu and Xia (2012), which is known to be computationally efficient.

The rest of this paper is organized as follows. In Section 2, we will introduce our new model selection method. Consistency of the model selection method is established in Section 3. In Section 4, we compare the performance of our proposal with that of previous methods developed for the same purpose in terms of model selection accuracy and computing time. All technical details are deferred to the Appendix.

2 A New Bayesian Information Criterion to Detect Constancy

In this section, we propose an information criterion that can consistently identify constant coefficients. For the estimation of model (1.2), suppose that we have a random sample of size n, $\{(Y_i, U_i, X_{1i}, \ldots, X_{pi}),\ i = 1, \ldots, n\}$. As a first step for model selection, we obtain the estimates $\hat{a}_l(U_i),\ i = 1, \ldots, n,\ l = 1, \ldots, p$, based on local linear regression. Using Taylor's expansion, we have

$$a_l(U) \cong a_l(u) + a_l'(u)(U - u), \quad l = 1, \ldots, p, \tag{2.1}$$

for U in a neighborhood of u. This leads to the following local least squares estimation problem:

$$(\tilde{\boldsymbol{b}}, \tilde{\boldsymbol{c}}) = \arg\min_{\boldsymbol{b},\boldsymbol{c}} \sum_{i=1}^{n} \left[Y_i - \sum_{l=1}^{p} \{b_l + c_l(U_i - u)\} X_{li} \right]^2 K_h(U_i - u), \tag{2.2}$$

with respect to $\boldsymbol{b} = (b_1, \ldots, b_p)^\top$ and $\boldsymbol{c} = (c_1, \ldots, c_p)^\top$, for a given kernel function K and a bandwidth h. Here, $K(\cdot)$ is a symmetric density function and $K_h(\cdot) = h^{-1}K(\cdot/h)$. Let $\tilde{a}_l(u)$ be $\tilde{b}_l$ for $l = 1, \ldots, p$ and let $\tilde{\boldsymbol{a}}(u) = (\tilde{a}_1(u), \ldots, \tilde{a}_p(u))^\top$.

Now, we are ready to define a Bayesian Information Criterion (BIC) to identify which coefficients are varying. Without loss of generality, we can assume that the whole set of covariates can be separated into two exclusive subsets $\mathcal{I}_v^0 = \{1, \ldots, s\}$ and $\mathcal{I}_c^0 = \{s+1, \ldots, p\}$ representing the indices of varying and constant coefficients, respectively. For any partition $\mathcal{I}_v \cup \mathcal{I}_c = \{1, \ldots, p\}$, we define

$$\hat{a}_l(U_i) \equiv \tilde{a}(U_i) \quad \text{for} \quad l \in \mathcal{I}_v \quad \text{and} \quad \hat{a}_l \equiv \frac{1}{n}\sum_{i=1}^{n} \tilde{a}(U_i) \quad \text{for} \quad l \in \mathcal{I}_c. \tag{2.3}$$

Based on these estimates, the BIC for detecting constancy is defined as follows:

$$\text{BIC}(\mathcal{I}_v, \mathcal{I}_c) = \log(\text{RSS}(\mathcal{I}_v, \mathcal{I}_c)) + \frac{\log(nh)}{nh}|\mathcal{I}_v| + \frac{\log n}{n}|\mathcal{I}_c|, \tag{2.4}$$

where

$$\text{RSS}(\mathcal{I}_v, \mathcal{I}_c) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{l\in\mathcal{I}_v}\hat{a}_l(U_i)X_{li} - \sum_{l\in\mathcal{I}_c}\hat{a}_l X_{li}\right)^2, \tag{2.5}$$

and $|\mathcal{I}|$ is the cardinality of the set $\mathcal{I}$. Since the set $\mathcal{I}_c$ is the complement of $\mathcal{I}_v$, we suppress $\mathcal{I}_c$ in the remaining of the paper. The set of indices of varying coefficient functions is estimated by the minimizer of $\text{BIC}(\mathcal{I}_v)$, which we denote by $\hat{\mathcal{I}}_v$. Similar criteria were considered in Hu and Xia (2012) and Cheng et al. (2009). Although Cheng et al. (2009) considered a similar criterion in a likelihood estimation framework, they did not focus on the consistency of the procedure to identify constant coefficients. Hu and Xia (2012) used another similar BIC to select the amount of shrinkage for the same identification problem. Different from theirs, in order to detect underfit more sensitively, we use the traditional residual sum of squares which does not involve smoothing, whereas Hu and Xia (2012) used the kernel-weighted residual sum of squares, which is common in kernel smoothing methods and whose theoretical properties are easy to show. In Section 4, we will show by means of simulations that this difference in the definition of $\text{RSS}(\mathcal{I}_v)$ is really crucial to enhance the ability of preventing underfit.

3 Consistency of the Model Selection Rule

To study consistency of the proposed method for model selection, the following standard regularity conditions are needed (Fan and Huang, 2005; Hu and Xia, 2012):

(A1) There is an $s > 2$ such that $E|\epsilon|^{2s} < \infty$, $E\|\boldsymbol{X}\|^{2s} < \infty$ and $n^{2\epsilon-1}h \to \infty$ for some $\epsilon < 2 - s^{-1}$, where $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ and $\|\boldsymbol{X}\|^2 = \boldsymbol{X}^\top\boldsymbol{X}$.

(A2) The support of the random variable U is $[0,1]$. The density function of U, $f(u)$, is Lipschitz continuous and bounded away from 0 on the support.

(A3) The $p \times p$ matrix $E(\boldsymbol{X}\boldsymbol{X}^\top|U=u)$ is nonsingular for each $u \in [0,1]$, and the functions $u \mapsto E(\boldsymbol{X}\boldsymbol{X}^\top|U=u)$ and $E(\boldsymbol{X}\boldsymbol{X}^\top|U=u)^{-1}$ are Lipschitz continuous. Moreover, the function $E(\|\boldsymbol{X}\|^4|U=u)$ is bounded.

(A4) The conditional density of U given $\boldsymbol{X}$ is continuous and uniformly bounded with respect to u and $\boldsymbol{x}$ up to its second derivative with respect to u.

(A5) The function $E(\epsilon^4|U=u,\boldsymbol{X}=\boldsymbol{x})$ and the second order derivative of $f(u)$ are bounded with respect to u and $\boldsymbol{x}$.

(A6) The second order derivatives of the coefficients $a_l(u),\ l=1,\ldots,p$, are continuous.

(A7) $K(u)$ is a symmetric density function with compact support.

(A8) $nh^5 \to \kappa$ with $0 < \kappa < \infty$ as $n \to \infty$.

We are now ready to show the consistency of the proposed BIC procedure.

Theorem 3.1 *Under Assumptions (A1)-(A8), we have*

$$\lim_{n\to\infty} \mathrm{P}(\hat{\mathcal{I}}_v = \mathcal{I}_v^0) = 1. \tag{3.1}$$

When the simple averaging approach in (2.3) for the estimation of constant coefficients is used as in Zhang et al. (2002), Lee (2003) and Cai and Xiao (2012), undersmoothing is inevitable for a root-n convergence rate of $\hat{a}_l$ ($l \in \mathcal{I}_c$). However, obtaining a root-n convergence rate for $\hat{a}_l$ ($l \in \mathcal{I}_c$) is not necessary for the consistency of our model selection method (as will be shown in the proof of Theorem 3.1). Since our theoretical result is established under the ordinary optimal bandwidth given in (A8), usual bandwidth selection methods based on the mean squared error can be used for our method. Further, the simple averaging approach enables us to choose the bandwidth only once, whereas the method of Xia et al. (2004) requires the selection of a bandwidth for each candidate model.

Note that one could consider another type of BIC for our method, which is directly motivated by Hu and Xia (2012):

$$\mathrm{BIC}^*(\mathcal{I}_v) = \log(\mathrm{RSS}^*(\mathcal{I}_v)) + \frac{\log(nh)}{nh}|\mathcal{I}_v| + \frac{\log n}{n}|\mathcal{I}_c|, \tag{3.2}$$

where

$$\mathrm{RSS}^*(\mathcal{I}_v) = \frac{1}{n^2}\sum_{t=1}^{n}\sum_{i=1}^{n}\left(Y_i - \sum_{l\in\mathcal{I}_v}\hat{a}_l(U_t)X_{li} - \sum_{l\in\mathcal{I}_c}\hat{a}_l X_{li}\right)^2 K_h(U_t - U_i). \tag{3.3}$$

Under assumptions similar to (A1)-(A8), it is possible to show that our method with $\mathrm{BIC}^*(\mathcal{I}_v)$ is also asymptotically consistent for the identification of constant coefficients. However, our simulation study suggests that $\mathrm{BIC}^*(\mathcal{I}_v)$ is significantly inferior to $\mathrm{BIC}(\mathcal{I}_v)$ in terms of resistance against underfit.

4 Numerical Studies

In this section we illustrate the competitiveness of our method in terms of model selection accuracy and computation time through the comparison with two previously proposed methods for the same purpose. Additionally, we illustrate the necessity to use $\mathrm{BIC}(\mathcal{I}_v)$ instead of $\mathrm{BIC}^*(\mathcal{I}_v)$ by means of simulations. Throughout this section, the kernel function $K(u) = \exp(-u^2/2)/\sqrt{2\pi}$ is used and the optimal bandwidth for our method is chosen by the leave-one-out cross-validation procedure.

4.1 Comparison with previous methods

We compare the finite sample performance of our method with two CV-based methods of Xia et al. (2004) and with the shrinkage method of Hu and Xia (2012). Concerning the CV methods, Xia et al. (2004) proposed two modifications of their original proposal. The first one is the simplified CV, which is designed to avoid an exhaustive search of the proposed CV criterion. The second one is an improvement of the first one in terms of overfit resistance. We included both improved versions for comparison, but not the original one because it is very time-consuming.

To compare the computation time of the different methods, we first implemented all methods in R software (R Core Team (2012)). In particular, as for the shrinkage method, we reimplemented it in R referring to its original implementation in MATLAB (2008) that Prof. Yingcun Xia kindly provided. However, we found that the implementation in MATLAB is somewhat faster than that in R, which is not the case for the other methods. Due to this, we report for the shrinkage method the computation time from its implementation in MATLAB.

We consider the following two sets of models:

(Model A)

$$Y_i = c(U_i - 0.5)^2 X_{1i} + X_{2i} + 0.5X_{3i} + 0.2\epsilon_i,$$

where $U_i \sim \text{Unif}[0,1]$, $X_{1i}, X_{2i}, X_{3i}, \epsilon_i \sim N(0,1)$ are all independent and the value of c controls the departure of the corresponding coefficient from a constant; and

(Model B)

$$\text{(B.1)} \qquad Y_i = 2\sin(2\pi U_i)X_{1i} + 4U_i(1-U_i)X_{2i} + 0X_{3i} + 0.5X_{4i} + 0.5X_{5i} + X_{6i} + 0.1X_{7i} + 0.5\epsilon_i\,;$$

$$\text{(B.2)} \quad Y_i = 3\sin(2\pi U_i)X_{1i} + 8U_i(1-U_i)X_{2i} + \cos^2(2\pi U_i)X_{3i} + X_{4i} + 0.5X_{5i} + X_{6i} - 0.5X_{7i} + 0.5\epsilon_i\,;$$

$$\text{(B.3)} \qquad Y_i = 3U_iX_{1i} + 2\sin(2\pi U_i)X_{2i} + 15U_i(1-U_i)X_{3i} + X_{4i} - X_{5i} + X_{6i} + 0X_{7i} + 0.5\epsilon_i,$$

where $U_i \sim \text{Unif}[0,1]$, $X_{1i} \equiv 1$, $(X_{2i}, \ldots, X_{7i})^\top$ is simulated from a multivariate normal with mean zero and $\text{cov}(X_{j_1 i}, X_{j_2 i}) = 0.5^{|j_1 - j_2|}$ for any $2 \le j_1, j_2 \le 7$, and ϵ_i follows a standard normal distribution. Models A and B were already considered in Xia et al. (2004) and Hu and Xia (2012), respectively. To evaluate the performance of the three model selection methods, we discriminate three different situations in Table 1 - 3. In column "C", we present the percentage of trials in which all the varying and constant coefficients are correctly identified. When the estimated model misses at least one varying coefficient, we count it as an underfitted model and report its percentage in column "U". Finally, we show in column "O" the percentage of overfitted cases in which all the varying coefficients are correctly identified but additionally some of the constant coefficients are identified as varying. We measure the computing time in seconds using the following computing environment with CPU: Intel(R) Core(TM) i7 2.80 GHz and RAM: 4GB.

Table 1 and 2 summarize the model selection results. The consistency rates for all methods come closer to 1 as the sample size grows. Only the simplified CV method seems to suffer a bit from slow convergence, as was already explained in Xia et al. (2004). In terms of model selection performance, both the modified CV method and ours work equally better than the other methods, but the former seems to be slightly better than the latter in Model A, which corresponds to the case where the non-constant coefficients do not strongly deviate from a constant. However, if we consider the computation time, our method appears to be incomparably better than all other methods including the shrinkage method, which is supposed to be computationally efficient. Finally, although the results are not presented here, we learned the same lesson for the case where different kinds of error distributions are used as long as the errors satisfy the assumptions in Section 3.

(*Table* 1)

(*Table* 2)

4.2 Comparison between $\mathrm{BIC}(\mathcal{I}_v)$ and $\mathrm{BIC}^*(\mathcal{I}_v)$

As will be shown in the proof of Theorem 3.1, the sum of residual squares in the BIC criterion prevents underfit, whereas the penalty resists against overfit. Our intuition in the proposal of $\mathrm{BIC}(\mathcal{I}_v)$ was that the sum of residual squares not involving smoothing would prevent underfitting more effectively than the one involving smoothing. To confirm this intuition, we compare $\mathrm{BIC}(\mathcal{I}_v)$ with $\mathrm{BIC}^*(\mathcal{I}_v)$ defined in (3.2) for Model A. Table 3 shows that $\mathrm{BIC}^*(\mathcal{I}_v)$ seriously underfits when $c = 1$, which confirms our intuition.

(*Table* 3)

5 Conclusion and Future Research

We developed a new model selection method for semivarying coefficient models, which is shown to be competitive in identification performance and significantly faster than previous methods studied in the literature. Although we only deal with the case of independent and identically distributed data, we expect that it would also work in a time series setting as the method of Xia et al. (2004) does. Such extension of our work would be an interesting topic for future research. Also, the extension of our work to quantile regression is worthy of research.

Acknowledgements

The authors would like to thank Prof. Yingcun Xia, who kindly provided the code of his method for our comparison. H. Noh and I. Van Keilegom acknowledge financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650. I. Van Keilegom also acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy) and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

References

Antoniadis, A., I. Gijbels, and A. Verhasselt (2012). Variable selection in varying-coefficient models using P-splines. *Journal of Computational and Graphical Statistics 21*, 638–661.

Cai, Z. and Z. Xiao (2012). Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics 167*, 413 – 425.

Cheng, M.-Y., W. Zhang, and L.-H. Chen (2009). Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association 104*, 1179–1191.

Fan, J. and T. Huang (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli 11*, 1031–1057.

Fan, J., C. Zhang, and J. Zhang (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics 29*, 153–193.

Hu, T. and Y. Xia (2012). Adaptive semi-varying coefficient model selection. *Statistica Sinica 22*, 575–599.

Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory 19*, 1–31.

MATLAB (2008). *version 7.6.0*. The MathWorks Inc.

Noh, H. and B. Park (2010). Sparse varying coefficient models for longitudinal data. *Statistica Sinica 20*, 1183–1202.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Wang, H. J., Z. Zhu, and J. Zhou (2009). Quantile regression in partially linear varying coefficient models. *Annals of Statistics 37*, 3841–3866.

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*, 1556–1569.

Xia, Y., W. Zhang, and H. Tong (2004). Efficient estimation for semivarying-coefficient models. *Biometrika 91*, 661–681.

Yao, F. and C. Martins-Filho (2012). An asymptotic characterization of finite order U-statistics with sample size dependent kernels. Technical report, West Virginia University.

Zhang, W., S. Lee, and X. Song (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis 82*, 166–188.

Appendix

Lemma 5.1 *(Yao and Martins-Filho, 2012, Theorem 1) Let* $\{\boldsymbol{Z}_i\}_{i=1}^n$ *be a sequence of independent and identically distributed random variables and* $\psi_n(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_k)$ *be a symmetric function with* $k \leq n$*. Let* u_n *denote a* k*-order* U*-statistic with kernel function* $\psi_n(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_k)$*, which is defined as*

$$u_n = \binom{n}{k}^{-1} \sum_{(n,k)} \psi_n(\boldsymbol{Z}_{i_1},\ldots,\boldsymbol{Z}_{i_k}), \tag{5.1}$$

where $\sum_{(n,k)}$ *denotes the sum over all subsets* $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ *of* $\{1,2,\ldots,n\}$*. In addition, define* $r_{cn}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_c) = E(\psi_n(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_c,\boldsymbol{Z}_{c+1},\ldots,\boldsymbol{Z}_k)|\boldsymbol{Z}_1 = \boldsymbol{z}_1,\ldots,\boldsymbol{Z}_c = \boldsymbol{z}_c)$*,* $\sigma_{cn}^2 = \mathrm{Var}(r_{cn}(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_c))$ *and* $\theta_n = E(\psi_n(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_k))$*. Then,*

$$u_n = \theta_n + \sum_{j=1}^{k} O_p((n^{-j}\sigma_{jn}^2)^{1/2}). \tag{5.2}$$

Further, for $1 \leq c_1 \leq c_2 \leq k$*, we have* $c_1^{-1}\sigma_{c_1 n}^2 \leq c_2^{-1}\sigma_{c_2 n}^2$*.*

Lemma 5.2 *Let* $\boldsymbol{a}(u) = (a_1(u),\ldots,a_p(u))^\top$ *and* $\tilde{\boldsymbol{a}}(u)$ *its local linear estimator defined as* $\tilde{\mathbf{b}}$ *in (2.2). The following results hold true for* $\tilde{\boldsymbol{a}}(u)$*,* $\hat{a}_l(u)$ *and* $\hat{a}_l$ *uniformly in* $u \in [0,1]$ *with* $c_n = O_p\left(h + (\log n/(nh))^{1/2}\right)$*:*

$$\begin{aligned}
\tilde{\boldsymbol{a}}(u) - \boldsymbol{a}(u) &= \left[\frac{1}{n}\sum_{i=1}^{n} f(u)\boldsymbol{\Gamma}^{-1}(u)\boldsymbol{X}_i \left\{\epsilon_i + \sum_{l=1}^{p}(a_l(U_i) - a_l^L(U_i;u))X_{li}\right\} K_h(U_i - u)\right] \\
&\quad \times(1 + O_p(c_n)); \qquad (5.3)\\
\tilde{a}_l(u) - a_l(u) &= O_p(c_n) \quad \forall l \in \mathcal{I}_v; \qquad (5.4)\\
\hat{a}_l &= \frac{1}{n}\sum_{i=1}^{n}\hat{a}_l(U_i) = \int_0^1 a_l(u)f(u)du + O_p(c_n) \quad \forall l = 1,\ldots,p; \qquad (5.5)\\
\hat{a}_l(u) - \hat{a}_l &= a_l(u) - \int_0^1 a_l(u)f(u)du + O_p(c_n) \quad \forall l \in \mathcal{I}_v; \qquad (5.6)\\
\hat{a}_l(u) - \hat{a}_l &= O_p(c_n) \quad \forall l \in \mathcal{I}_c, \qquad (5.7)
\end{aligned}$$

where $\boldsymbol{\Gamma}(u) = E(\boldsymbol{X}\boldsymbol{X}^\top|U = u)$ *and* $a_l^L(x;u) = a_l(u) + a_l'(u)(x-u)$ *is the linear approximation of* $a_l(x)$ *at* u*.*

Proof. The result can be easily shown using similar arguments as in the Appendix of Fan and Huang (2005).

Proof of Theorem 3.1. To avoid digression, we suppose throughout that the true model is given by

$$Y_i = a_1(U_i)X_{1i} + a_2X_{2i} + a_3X_{3i} + \epsilon_i. \tag{5.8}$$

In this case, note that $\mathcal{I}_v^0 = \{1\}$, $\mathcal{I}_c^0 = \{2,3\}$, and $a_2(x) - a_2^L(x;u) = a_3(x) - a_3^L(x;u) = 0$.

Case 1. (Underfitted Model) We consider the situation where $\mathcal{I}_v = \{3\}$ as a generic case of the underfitted models. To show that our method is underfit-resistant, it is enough to show that there exists a $C > 0$ such that $P(\mathrm{RSS}(\mathcal{I}_v) - \mathrm{RSS}(\mathcal{I}_v^0) \geq C) \to 1$ as $n \to \infty$. This follows from the fact that

$$\begin{aligned}
\mathrm{RSS}(\mathcal{I}_v) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{a}_1X_{1i} - \hat{a}_2X_{2i} - \hat{a}_3(U_i)X_{3i})^2 \\
&= \mathrm{RSS}(\mathcal{I}_v^0) + \frac{1}{n}\sum_{i=1}^{n}\{(\hat{a}_1(U_i) - \hat{a}_1)^2X_{1i}^2 + (\hat{a}_3(U_i) - \hat{a}_3)^2X_{3i}^2\} \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\epsilon_i^*(\hat{a}_1(U_i) - \hat{a}_1)X_{1i} - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i^*(\hat{a}_3(U_i) - \hat{a}_3)X_{3i} \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}(\hat{a}_1(U_i) - \hat{a}_1)X_{1i}(\hat{a}_3(U_i) - \hat{a}_3)X_{3i} \\
&= \mathrm{RSS}(\mathcal{I}_v^0) + E\{(a_1(U) - Ea_1(U))^2X_1^2\} + O_p(c_n),
\end{aligned}$$

where $\epsilon_i^* = Y_i - \hat{a}_1(U_i) - \hat{a}_2X_{2i} - \hat{a}_3X_{3i}$.

Case 2. (Overfitted Model) We only consider the situation where $\mathcal{I}_v = \{1,2\}$. For the other cases similar arguments can be used to show that any overfitted model will not be chosen by the BIC with probability tending to 1. To show that asymptotically the BIC does not select $\mathcal{I}_v = \{1,2\}$, we only need to prove that

$$\mathrm{RSS}(\mathcal{I}_v) - \mathrm{RSS}(\mathcal{I}_v^0) = o_p\left(\frac{\log(nh)}{nh}\right). \tag{5.9}$$

Note that $\mathrm{RSS}(\mathcal{I}_v^0) \to \sigma^2 = E(\epsilon^2)$ as $n \to \infty$. Therefore, once (5.9) is shown, we have

$$\begin{aligned}
\mathrm{BIC}(\mathcal{I}_v) - \mathrm{BIC}(\mathcal{I}_v^0) &= \log\left(1 + \frac{\mathrm{RSS}(\mathcal{I}_v) - \mathrm{RSS}(\mathcal{I}_v^0)}{\mathrm{RSS}(\mathcal{I}_v^0)}\right) + \left(\frac{\log(nh)}{nh} - \frac{\log n}{n}\right)(|\mathcal{I}_v| - |\mathcal{I}_v^0|) \\
&\geq o_p\left(\frac{\log(nh)}{nh}\right) + \frac{\log(nh)}{nh}(1 + o_p(1)),
\end{aligned}$$

which shows that $P(\mathrm{BIC}(\mathcal{I}_v) - \mathrm{BIC}(\mathcal{I}_v^0) > 0) \to 1$ as $n \to \infty$. To show (5.9), we consider the following decomposition:

$$\begin{aligned}
& \mathrm{RSS}(\mathcal{I}_v) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{a}_1(U_i)X_{1i} - \hat{a}_2(U_i)X_{2i} - \hat{a}_3 X_{3i})^2 \\
= \; & \mathrm{RSS}(\mathcal{I}_v) - \frac{2}{n}\sum_{i=1}^{n}(Y_i - \hat{a}_1(U_i)X_{1i} - \hat{a}_2 X_{2i} - \hat{a}_3 X_{3i})(\hat{a}_2(U_i) - \hat{a}_2)X_{2i} + \frac{1}{n}\sum_{i=1}^{n}(\hat{a}_2(U_i) - \hat{a}_2)^2 X_{2i} \\
\equiv \; & \mathrm{RSS}(\mathcal{I}_v^0) - 2A + B. \qquad (5.10)
\end{aligned}$$

The decomposition reduces the problem to showing that both A and B are $o_p\left(\log(nh)/nh\right)$, which is given below. We use the abbreviated notations $\mathbf{\Gamma}_i$, f_i, $\boldsymbol{Z}_i$ and K_{ij} for $\mathbf{\Gamma}(U_i)$, $f(U_i)$, $(\boldsymbol{X}_i, U_i, \epsilon_i)^\top$ and $K_h(U_i - U_j)$, respectively. Additionally, we use $a_n \approx b_n$ when $a_n/b_n = O_p(1)$ and $b_n/a_n = O_p(1)$ and use $\boldsymbol{e}_i$ to denote the ith standard basis vector of $\mathbb{R}^p$. To show that $A = o_p\left(\log(nh)/nh\right)$, observe that

$$\begin{aligned}
A \; = \; & \frac{1}{n}\sum_{i=1}^{n}\left\{\epsilon_i - (\hat{a}_1(U_i) - a_1(U_i))X_{1i} - (\hat{a}_2 - a_2)X_{2i} - (\hat{a}_3 - a_3)X_{3i}\right\} \times \left\{\hat{a}_2(U_i) - a_2 - (\hat{a}_2 - a_2)\right\} X_{2i} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\hat{a}_2(U_i) - a_2)X_{2i} - \frac{1}{n}\sum_{i=1}^{n}(\hat{a}_1(U_i) - a_1(U_i))X_{1i}(\hat{a}_2(U_i) - a_2)X_{2i} \\
& - \frac{1}{n}\sum_{i=1}^{n}\left\{(\hat{a}_2 - a_2)X_{2i} + (\hat{a}_3 - a_3)X_{3i}\right\}(\hat{a}_2(U_i) - a_2)X_{2i} \\
& - \frac{1}{n}\sum_{i=1}^{n}\left\{\epsilon_i - (\hat{a}_1(U_i) - a_1(U_i))X_{1i} - (\hat{a}_2 - a_2)X_{2i} - (\hat{a}_3 - a_3)X_{3i}\right\}(\hat{a}_2 - a_2)X_{2i} \\
\equiv \; & A_1 - A_2 - A_3 - A_4.
\end{aligned}$$

As for the first term A_1, from (5.3) we have that

$$\begin{aligned}
A_1 \; \approx \; & \frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{2i}\boldsymbol{e}_2^\top\left(\frac{1}{n}\sum_{j=1}^{n}f_i^{-1}\mathbf{\Gamma}_i^{-1}\boldsymbol{X}_j(\epsilon_j + a_{ji}^*)K_{ji}\right) \\
= \; & \frac{1}{n^2}\sum_{i\neq j}\boldsymbol{e}_2^\top\mathbf{\Gamma}_i^{-1}\boldsymbol{X}_j X_{2i}K_{ji}\epsilon_i(\epsilon_j + a_{ji}^*)f_i^{-1} + \frac{1}{n^2 h}\sum_{i=1}^{n}\boldsymbol{e}_2^\top\mathbf{\Gamma}_i^{-1}\boldsymbol{X}_i X_{2i}K(0)\epsilon_i^2 f_i^{-1} \\
\equiv \; & \frac{1}{n^2}\sum_{i\neq j}\psi_{nij} + O_p\left(\frac{1}{nh}\right) = \frac{1}{2}\left(O(n^{-3}) + \binom{n}{2}^{-1}\right)\sum_{i<j}(\psi_{nij} + \psi_{nji}) + O_p\left(\frac{1}{nh}\right) \\
\equiv \; & \frac{1}{2}\left(O(n^{-3}) + \binom{n}{2}^{-1}\right)\sum_{i<j}\phi_{nij} + O_p\left(\frac{1}{nh}\right),
\end{aligned}$$

where $a_{ij}^* = \sum_{l=1}^{3}\{a_l(U_i) - a_l^L(U_i;U_j)\}X_{li} = (a_1(U_i) - a_1^L(U_i;U_j))X_{1i}$. Here, $u_n^1 \equiv \binom{n}{2}^{-1}\sum_{i<j}\phi_{nij}$ is a second order U-statistic. Since $\theta_n = E(\phi_{nij}) = 0$, $\sigma_{1n}^2 = \mathrm{Var}\{E(\phi_{nij}|\boldsymbol{Z}_i)\} \leq 2E\{E^2(\psi_{nij}|\boldsymbol{Z}_i) +$

$E^2(\psi_{nij}|\boldsymbol{Z}_j)\} = O(h^4)$ and $\sigma_{2n}^2 = \mathrm{Var}(\phi_{nij}) \leq E(\phi_{nij}^2) = O(h^{-1})$, Lemma 5.1 implies that $u_n^1 = O_p(n^{-1/2}h^2 + n^{-1}h^{-1/2})$, which results in $A_1 = o_p\left(\log(nh)/nh\right)$.

Regarding the second term A_2, first observe that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(\hat{a}_1(U_i) - a_1(U_i))X_{1i}(\hat{a}_2(U_i) - a_2)X_{2i} \\
\approx\ & \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{e}_1^T\left(\frac{1}{n}\sum_{j=1}^{n} f_i^{-1}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_j(\epsilon_j + a_{ji}^*)K_{ji}\right)X_{1i} \times \boldsymbol{e}_2^\top\left(\frac{1}{n}\sum_{k=1}^{n} f_i^{-1}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_k(\epsilon_k + a_{ki}^*)K_{ki}\right)X_{2i}\right\} \\
=\ & \frac{1}{n^3}\sum_{i,j,k}\boldsymbol{e}_1^\top\boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_j\boldsymbol{e}_2^\top\boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_k X_{1i}X_{2i}(\epsilon_j + a_{ji}^*)(\epsilon_k + a_{ki}^*)K_{ji}K_{ki}f_i^{-2} \equiv \frac{1}{n^3}\sum_{i,j,k}\psi_{nijk} \\
=\ & \frac{1}{n^3}\sum_{i=j\neq k}\psi_{nijk} + \frac{1}{n^3}\sum_{i=k\neq j}\psi_{nijk} + \frac{1}{n^3}\sum_{i\neq j, j\neq k, i\neq k}\psi_{nijk} + \frac{1}{n^3}\sum_{j=k\neq i}\psi_{nijk} + \frac{1}{n^3}\sum_{i=j=k}\psi_{nijk} \\
\equiv\ & A_{2,1} + A_{2,2} + A_{2,3} + A_{2,4} + A_{2,5}.
\end{aligned}
$$

It is obvious that $A_{2,5} = O_p(1/(nh)^2)$. Here we will focus only on $A_{2,3}$ and $A_{2,4}$ but the terms $A_{2,1}$ and $A_{2,2}$ can be shown to have the desired order using similar arguments as the ones for $A_{2,3}$ and $A_{2,4}$. To show that $A_{2,3} = o_p\left(\log(nh)/nh\right)$, observe that $A_{2,3}$ is equivalent to the following third order U-statistic $u_n^{2,3}$:

$$
A_{2,3} \approx u_n^{2,3} \equiv \frac{1}{\binom{n}{3}}\sum_{i<j<k}(\psi_{nijk} + \psi_{nikj} + \psi_{njik} + \psi_{njki} + \psi_{nkij} + \psi_{nkji}) \equiv \frac{1}{\binom{n}{3}}\sum_{i<j<k}\phi_{nijk}
$$

Note that $\theta_n = E(\phi_{nijk}) = O(h^4)$, $\sigma_{3n}^2 = \mathrm{Var}(\phi_{nijk}) \leq E(\phi_{nijk}^2) = O(h^{-2})$ and $\sigma_{2n}^2 = \mathrm{Var}\{E(\phi_{nijk}|\boldsymbol{Z}_i)\} \leq (2/3)\sigma_{3n}^2$. Further, by standard calculations in kernel smoothing, it can be shown that $E(E^2(\psi_{nijk}|\boldsymbol{Z}_i)) = O(h^8)$, $E(E^2(\psi_{nijk}|\boldsymbol{Z}_j)) = 0$ and $E(E^2(\psi_{nijk}|\boldsymbol{Z}_k)) = 0$, because $E(\psi_{nijk}|\boldsymbol{Z}_j, \boldsymbol{Z}_k) = 0$ from (A7) when $j \neq k$. Since $\sigma_{1n}^2 = \mathrm{Var}\{E(\phi_{nijk}|\boldsymbol{Z}_i, \boldsymbol{Z}_j)\} \leq 3E[E^2(\psi_{nijk}|\boldsymbol{Z}_i) + E^2(\psi_{nijk}|\boldsymbol{Z}_j) + E^2(\psi_{nijk}|\boldsymbol{Z}_k)]$, we have that $\sigma_{1n}^2 = O(h^8)$ and hence $u_n^{2,3} = O_p(h^4 + n^{-1/2}h^4 + n^{-1}h^{-1} + n^{-3/2}h^{-1}) = o_p(\log(nh)/nh)$ by Lemma 5.1. As for $A_{2,4}$, note that

$$
\begin{aligned}
A_{2,4} &\approx \frac{1}{n}\frac{1}{\binom{n}{2}}\sum_{i\neq k}\boldsymbol{e}_1^\top\boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_k\boldsymbol{e}_2^\top\boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_k X_{1i}X_{2i}(\epsilon_k + a_{ki}^*)^2K_{ki}^2f_i^{-2} \\
&\equiv \frac{1}{n}\frac{1}{\binom{n}{2}}\sum_{i\neq k}\psi_{nik} = \frac{1}{n}\frac{1}{\binom{n}{2}}\sum_{i<k}(\psi_{nik} + \psi_{nki}) \equiv \frac{1}{n}\frac{1}{\binom{n}{2}}\sum_{i<k}\phi_{nik} \equiv \frac{1}{n}u_n^{2,4}.
\end{aligned}
$$

Since $2\sigma_{1n}^2 \leq \sigma_{2n}^2 \leq E(\phi_{nik}^2) = O(h^{-3})$ and $\theta_n = E(\phi_{nik}) = O(h^{-1})$, we have $u_n^{2,4} = O_p(h^{-1} + n^{-1/2}h^{-3/2} + n^{-1}h^{-3/2})$ by Lemma 5.1, and hence $A_{2,4} = o_p\left(\log(nh)/nh\right)$.

Using similar calculations to those for A_1 and A_2, it can be easily shown that

$$
\hat{a}_2 - a_2 = \hat{a}_3 - a_3 = O_p(h^2 + n^{-1/2}). \tag{5.11}
$$

One can show that both A_3 and A_4 are $o_p\left(\log(nh)/nh\right)$ using (5.11) and U-statistic computations similar to those done for A_1 and A_2. This shows that $A = o_p\left(\log(nh)/nh\right)$.

Let us now consider the term B in (5.10). Consider the decomposition

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}(\hat{a}_2(U_i) - \hat{a}_2)^2 X_{2i}^2 \\ = \; & \frac{1}{n}\sum_{i=1}^{n}(\hat{a}_2(U_i) - a_2)^2 X_{2i}^2 + (\hat{a}_2 - a_2)^2 \frac{1}{n}\sum_{i=1}^{n} X_{2i}^2 - (\hat{a}_2 - a_2)\frac{2}{n}\sum_{i=1}^{n}(\hat{a}_2(U_i) - a_2)X_{2i}^2. \quad (5.12) \end{aligned}$$

Regarding the first two terms, it is easy to show that they have the desired order using similar arguments as for the term A and the third term in (5.12). Therefore, we concentrate on the third term. Observe that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(\hat{a}_2(U_i) - a_2)X_{2i}^2 &\approx \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_{2i}^2 \boldsymbol{e}_2^\top \boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_j(\epsilon_j + a_{ji}^*)K_{ji}f_i^{-1} \\ &= \frac{1}{2}\left(O(n^{-3}) + \binom{n}{2}^{-1}\right)\sum_{i\neq j} X_{2i}^2 \boldsymbol{e}_2^\top \boldsymbol{\Gamma}_i^{-1}\boldsymbol{X}_j(\epsilon_j + a_{ji}^*)K_{ji}f_i^{-1} + O_p\left(\frac{1}{nh}\right) \\ &\equiv \frac{1}{2}\left(O(n^{-3}) + \binom{n}{2}^{-1}\right)\sum_{i\neq j}\psi_{nij} + O_p\left(\frac{1}{nh}\right) \\ &\equiv \frac{1}{2}\left(O(n^{-3}) + \binom{n}{2}^{-1}\right)\sum_{i<j}\phi_{nij} + O_p\left(\frac{1}{nh}\right), \end{aligned}$$

where $\phi_{nij} = \psi_{nij} + \psi_{nji}$. Since $u_n^B \equiv \binom{n}{2}^{-1}\sum_{i<j}\phi_{nij}$ is a second order U-statistic, by similar arguments as before we can easily show that $u_n^B = O_p(h^2 + n^{-1/2}h^{-1/2})$, which means that the third term in (5.12) is $o_p\left(\log(nh)/nh\right)$ because of (5.11).

Table 1: Comparison of the results of the two CV methods in Xia et al. (2004), our proposal and the shrinkage method of Hu and Xia (2012). The results are based on 400 data sets generated under Model A.

		Simplified CV				Modified CV				BIC				Shrinkage			
c	n	C	U	O	Time	C	U	O	Time	C	U	O	Time	C	U	O	Time
1	50	0.50	0.30	0.20	22.51	0.58	0.36	0.06	22.52	0.51	0.35	0.14	5.98	0.02	0.98	0.00	11.88
	100	0.75	0.09	0.16	54.63	0.85	0.13	0.02	56.32	0.74	0.24	0.02	14.12	0.03	0.97	0.00	40.65
	200	0.83	0.00	0.17	164.16	0.96	0.01	0.03	221.43	0.93	0.07	0.00	40.25	0.05	0.95	0.00	164.58
2	50	0.76	0.02	0.22	22.12	0.88	0.06	0.06	22.19	0.79	0.01	0.20	5.85	0.33	0.66	0.01	13.08
	100	0.85	0.00	0.15	49.24	0.98	0.00	0.02	56.30	0.98	0.00	0.02	12.69	0.66	0.34	0.00	43.12
	200	0.86	0.00	0.14	149.31	0.99	0.00	0.01	221.80	1.00	0.00	0.00	36.56	0.98	0.02	0.00	158.65

Table 2: Comparison of the results of the two CV methods in Xia et al. (2004), our proposal and the shrinkage method of Hu and Xia (2012). The results are based on 100 data sets generated under Model B.

		Simplified CV				Modified CV				BIC				Shrinkage			
Model	n	C	U	O	Time	C	U	O	Time	C	U	O	Time	C	U	O	Time
B.1	100	0.64	0.00	0.36	120.81	0.65	0.00	0.35	163.57	0.68	0.00	0.32	14.05	0.26	0.71	0.03	39.26
	200	0.63	0.00	0.37	396.52	1.00	0.00	0.00	674.37	0.96	0.00	0.04	45.48	0.39	0.59	0.03	135.39
	400	0.73	0.00	0.27	2821.69	1.00	0.00	0.00	3064.92	1.00	0.00	0.00	309.48	0.87	0.10	0.03	1476.83
B.2	100	0.62	0.00	0.38	83.82	0.48	0.00	0.52	155.24	0.66	0.00	0.34	11.69	0.25	0.71	0.04	42.14
	200	0.86	0.00	0.14	344.01	0.94	0.00	0.06	605.95	0.95	0.00	0.02	42.41	0.42	0.50	0.08	155.19
	400	0.83	0.00	0.17	2492.94	1.00	0.00	0.00	2777.70	1.00	0.00	0.00	305.20	0.85	0.13	0.02	2112.84
B.3	100	0.59	0.00	0.41	105.22	1.00	0.00	0.00	143.62	0.79	0.00	0.21	14.85	0.66	0.00	0.34	46.51
	200	0.70	0.00	0.30	418.60	1.00	0.00	0.00	602.31	0.98	0.00	0.02	55.20	0.90	0.00	0.10	252.89
	400	0.82	0.00	0.18	2568.24	1.00	0.00	0.00	2769.74	1.00	0.00	0.00	307.80	1.00	0.00	0.00	1675.25

Table 3: Comparison of the results of two versions of our method. The results are based on 400 data sets generated under Model A.

		BIC($\mathcal{I}_v$)			BIC*($\mathcal{I}_v$)		
c	n	C	U	O	C	U	O
1	50	0.51	0.35	0.14	0.01	0.99	0.00
	100	0.74	0.24	0.02	0.01	0.99	0.00
	200	0.93	0.07	0.00	0.02	0.98	0.00
2	50	0.79	0.01	0.20	0.14	0.86	0.00
	100	0.98	0.00	0.02	0.43	0.57	0.00
	200	1.00	0.00	0.00	0.95	0.05	0.00