Electronic Journal of Statistics
Vol. 10 (2016) 855–894
ISSN: 1935-7524
DOI: 10.1214/16-EJS1110

Local linear smoothing for sparse high dimensional varying coefficient models

Eun Ryung Lee*

Department of Statistics, Sungkyunkwan University, 32419 Seoul, Korea
e-mail: silverryuee@gmail.com

and

Enno Mammen†

Institute for Applied Mathematics, Heidelberg University, 69120 Heidelberg, Germany
National Research University Higher School of Economics, 103012 Moscow, Russian Federation
e-mail: mammen@math.uni-heidelberg.de

Abstract: Varying coefficient models are useful generalizations of parametric linear models. They allow for parameters that depend on a covariate or that develop in time. They have a wide range of applications in time series analysis and regression. In time series analysis they have turned out to be a powerful approach to infer on behavioral and structural changes over time. In this paper, we are concerned with high dimensional varying coefficient models including the time varying coefficient model. Most studies in high dimensional nonparametric models treat penalization of series estimators. On the other side, kernel smoothing is a well established, well understood and successful approach in nonparametric estimation, in particular in the time varying coefficient model. But not much has been done for kernel smoothing in high-dimensional models. In this paper we will close this gap and we develop a penalized kernel smoothing approach for sparse high-dimensional models. The proposed estimators make use of a novel penalization scheme working with kernel smoothing. We establish a general and systematic theoretical analysis in high dimensions. This complements recent alternative approaches that are based on basis approximations and that allow more direct arguments to carry over insights from high-dimensional linear models. Furthermore, we develop theory not only for regression with independent observations but also for local stationary time series in high-dimensional sparse varying coefficient models. The development of theory for local stationary processes in a high-dimensional setting creates technical challenges. We also address issues of numerical implementation and of data adaptive selection of tuning parameters for penalization.The finite sample performance of the proposed methods is studied by simulations and it is illustrated by an empirical analysis of NASDAQ composite index data.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35.

*The financial support from the Research Center (SFB) 884 Political Economy of Reform (Project B6), funded by the German Research Foundation (DFG) is acknowledged.

†The financial support by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953 and from the Government of the Russian Federation within the framework of the implementation of the Global Competitiveness Program of the National Research University Higher School of Economics are gratefully acknowledged.

Keywords and phrases: Sparse estimation, local stationary time series, high-dimensional data, local stationarity, time varying coefficient models, kernel method, local linear method, penalized methods, BIC, oracle inequality, oracle property, partially linear varying coefficient model, semiparametric model, consistent structural identification, second order cone programming.

Received July 2015.

Contents

1. Introduction

Varying coefficient models arise in a wide range of applications. They are an important generalization of parametric linear regression models. They relax the assumption that the parameters are constant and allow regression coefficients to be smooth functions of other predictors, called index variables. On the one side, the models are very flexible and give an accurate fit of complex data and on the other side they still maintain a simple structure. This allows an intuitive interpretation and an accurate estimation. For an overview on varying coefficient models, we refer to Fan and Zhang [13] and Park et al. [26].

In this paper we will propose an approach based on kernel smoothing for sparse high-dimensional varying coefficient models. Kernel smoothing has yet

been considered mostly only for finite dimensional models. This is the case for varying coefficient models and for other nonparametric settings. Typically, work on sparse nonparametric high-dimensional models has made use of orthogonal series estimators. These estimators are more closely linked to linear models and for this reason they more easily allow to carry over theory from high-dimensional linear models. Our paper argues that also for high-dimensional nonparametric models kernel smoothing is an attractive alternative to orthogonal series estimation. This will be shown for varying coefficient models. Our implementation of kernel smoothing in high-dimensional settings requires the introduction of novel penalization schemes. We will show that kernel smoothing inherits from finite-dimensional models its intuitive interpretation and clear asymptotic theory for the distribution of the estimator.

In our theory we will consider both, regression models and time series models. In time series a central example of a varying coefficient model is the time varying coefficient model where the index variable is rescaled time. This class of models has been developed independently from varying coefficient models and it has turned out to be a very powerful tool in the empirical analysis for structural changes over time in time series data [see 27, 28, 9, 10, 2, 3, 44, for example]. An important example in this class is the time varying autoregressive model. In this model the data are non stationary because the autoregressive structure changes over time. This complicates the asymptotic analysis. A common strategy to handle this nonstationarity is to model the time series as locally stationary processes, see Dahlhaus [8, 9]. Roughly speaking, a locally stationary process behaves approximately as a stationary process over a short period of time. This naturally suggests the use of local smoothing methods like kernel smoothing [see 32, for example]. Estimation and statistical inference based on kernel smoothing has been established and their statistical properties have been well understood in the time varying coefficient model [2, 3, 44]. However, all this work is restricted to finite dimensional settings. As noted in Fan et al. [14], high dimensionality is encountered in many time series data applications, e.g. in economics and finance. Besides exogenous variables, often lagged variables of different lag orders and interaction terms have to be included into the model for accurate fits. These applications serve as an important motivation for our paper.

Sparse modeling provides an effective framework to analyze high dimensional data. It allows for identifiability of the model and it facilitates consistent statistical estimation even in high dimensional situations. Many penalized methods such as Least Absolute Shrinkage and Selection Operator [LASSO, 29] and Smoothly Clipped Absolute Deviation [SCAD, 12] have been proposed for variable selection and estimation in sparse linear regression. The methods have proven to possess high computational efficiency as well as desirable statistical properties even under high dimensional settings. This has motivated to extend the ideas to varying coefficient models for i.i.d. and longitudinal data. Varying coefficient models using orthogonal series estimation have been considered in Wei et al. [38], Lian [24], Xue and Qu [40] and Klopp and Pensky [21]. Their asymptotics allowed for an increasing number of coefficients and the studies include variable selection based on groupwise penalized methods such as the group

LASSO [41]. Moreover, Klopp and Pensky [21] developed a non-asymptotic minimax theory for a model where the coefficient functions possibly have different degrees of smoothness and where they are spatially inhomogeneous. All these papers do not treat kernel smoothing nor time series models. Furthermore, the theoretical studies heavily rely on the assumption of independent observations and partially they need that the covariables $\mathbf{X}_i$ and the predictor Z_i are independent, see (2.1) for the definition of $\mathbf{X}_i, Z_i$. This could be considered as a restrictive assumption. We will drop this condition on the way to cover time series models. For an initial screening procedure to handle ultra-high dimensional variables see also Cheng et al. [5], Fan et al. [15] and Cheng et al. [6]. But, not much is known on penalized kernel smoothing methods. For varying coefficient models, the only work we are aware of are Wang and Xia [35], Hu and Xia [19], Wang and Kulasekera [33] and Kong et al. [22]. However, their asymptotic analysis is restricted to the case of fixed-dimension and only the case of independent observations is treated.

Kernel smoothing is a very popular estimation technique for a lot of nonparametric models and it is especially recommended to use for the time varying coefficient models. In this paper we will develop kernel smoothing techniques that are working theoretically and computationally for varying coefficient models with a diverging number of variables. Our first contribution to accomplish this task is to propose a penalized local linear kernel estimation method in varying coefficient models and to provide its sound asymptotic theory under high dimensionality. We will adapt the group LASSO and SCAD methods to the local linear method and we will systematically study variable selection and estimation properties of these methods. Our theory will include oracle inequalities of the group LASSO kernel method and we will show that the group SCAD kernel method consistently identifies the true structure of a partially linear varying coefficient model. Our methodological and theoretical developments require technical treatments that are quite different from asymptotics for groupwise penalized methods using series estimators. For example, in the sieve approach, one approximates a nonparametric model by a parametric model with increasing dimension. Thus the estimation problem of the nonparametric model is methodologically very similar to the estimation of a parametric model with increasing dimension. Such a simplifying technical approach does not apply to kernel smoothing. Furthermore, we also treat local stationary varying coefficient models including the above-mentioned time varying autoregressive model. The study of this class of models requires new mathematical tools. Locally in time the time series has to be approximated by a stationary process. This approximation facilitates to carry over techniques from the study of stationary processes. We are not aware that such a theoretical study has been done in another high dimensional nonparametric set up. Our theory includes models with errors that have serial correlations with lagged errors and observations and with covariates. In particular, we allow for conditional heteroskedastic errors. We also do not assume that the errors are sub-Gaussian. The latter point may be important in financial applications.

Our second contribution is to develop a new computation method for the implementation of our proposals. Implementing our estimator involves a quite complicated optimization problem to which a typical group LASSO algorithm cannot be applied. By reformulating the problem as a second order cone programming problem, we are able to provide a simple and computationally efficient algorithm for the implementation. The details can be found in Section 4.1. The third contribution is to develop a criterion for determining the amount of penalization in the penalized estimation. This is a crucial step in the identification of the true partially linear structure. Although penalization methods for consistent identification of semiparametric models have been proposed, see Cheng et al. [5] and Zhang et al. [43], to our knowledge no work has been done on the choice of the tuning parameters for such estimators. We propose a tuning parameter selector based on the Bayesian information criterion (BIC) and we provide its theoretical justification. For this task, we verify that our penalized estimators of the relevant parametric and nonparametric components achieve the respective optimal rates of convergence at the same time. The result is new and, compared to the usual oracle properties in the literature (see Theorem 1 and Remark 3 in Zhang et al. [43] and Theorem 3.3 in Cheng et al. [5]) it is much stronger. It will be used as our theoretical foundation that the proposed BIC identifies the true partially linear structure with probability tending to one. Finally, even in the fixed dimensional case, our methods extend other kernel smoothing-based penalization methods.

The rest of this paper is organized as follows. The next section introduces the model and our statistical procedures based on kernel smoothing: LASSO-estimators that uses L_1-penalties for non-zero and non-linear component functions, SCAD-estimators with a BIC-choice of their penalty constants and BIC-choices of the set of non-zero coefficient function and of the set of non-linear coefficient functions. Section 3 contains our theoretical results. Section 4 discusses the numerical implementation of our methods and shows some simulation results. An illustrative data example is given in Section 5. All proofs are deferred to Appendix.

2. Model and methodology

We suppose that the data $\{(\mathbf{X}_i, Z_i, Y_i), 1 \leq i \leq n\}$ are generated under the model

$$Y_i = \sum_{j=1}^{p} X_i^{(j)} m_j^0(Z_i) + \epsilon_i, \tag{2.1}$$

where $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(p)})^\top$ are p-dimensional vectors, $Z_i \in [0,1]$ and ϵ_i are random errors. With rescaled time $Z_i = i/n$ we get the so-called 'time varying regression model'. This model includes the time varying autoregressive model as a special case. In our paper, we consider both, independent data and time series versions of model (2.1): in the i.i.d. scenario, we assume that $(\mathbf{X}_i, Z_i, \epsilon_i)$ in (2.1) are independent and identically distributed (i.i.d.) copies of $(\mathbf{X}, Z, \epsilon)$

with $E(\epsilon|\mathbf{X}, Z) = 0$ and $E(\epsilon^2|\mathbf{X}, Z) \leq \sigma^2 < \infty$ for some $\sigma^2 > 0$; in the time series scenario, we suppose that $Z_i = i/n$, $E(\epsilon_i|\ \mathbf{X}_t, \epsilon_{t-1},\ t \leq i) = 0$ and $E(\epsilon_i^2|\ \mathbf{X}_t, \epsilon_{t-1},\ t \leq i) \leq \sigma^2 < \infty$. We allow p to tend to infinity as $n \to \infty$. Our main assumption is the sparsity of the model (2.1), that is, $m_j^0 \equiv 0$ for many j's, as specified in more detail below.

Let $(m_j^0)^{(s)}$ be the s-th derivative of the true coefficient function m_j^0 for $1 \leq j \leq p$. Given any $z \in [0,1]$, we define $\mathbf{m}^0(z) = (m_1^0(z), \ldots, m_p^0(z))^\top$ and $(\mathbf{m}^0)^{(1)}(z) = ((m_1^0)^{(1)}(z), \ldots, (m_p^0)^{(1)}(z))^\top$. Motivated by a local approximation of $m_j^0(Z)$, $m_j^0(Z) \approx m_j^0(z) + (m_j^0)^{(1)}(z)(Z - z)$, for $Z \approx z$, the local linear estimator of $\mathbf{m}^0(z)$ and $(\mathbf{m}^0)^{(1)}(z)$, $z \in [0,1]$ is defined by minimizing the following local kernel weighted least squares criterion:

$$\begin{pmatrix} \bar{\mathbf{m}}(z) \\ \bar{\mathbf{m}}^{(1)}(z) \end{pmatrix} = \underset{\mathbf{m},\, \mathbf{m}^{(1)} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \Big[Y_i - \mathbf{X}_i^\top (\mathbf{m} + \mathbf{m}^{(1)}(Z_i - z)) \Big]^2 K_h(z - Z_i),$$

where $K_h(u) = K(u/h)/h$, K is a kernel function and $h > 0$ is a bandwidth. Equivalently, the estimated coefficient functions $\bar{\mathbf{m}}(\cdot)$ and $\bar{\mathbf{m}}^{(1)}(\cdot)$ are the minimizer of

$$L(\mathbf{m}, \mathbf{m}^{(1)}) = n^{-1} \int \sum_{i=1}^{n} \Big[Y_i - \mathbf{X}_i^\top (\mathbf{m}(z) + \mathbf{m}^{(1)}(z)(Z_i - z)) \Big]^2 K_h(z - Z_i) dz$$

with respect to $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$. From now on, we omit the arguments of functions when no confusion arises.

Given a function g defined on $[0,1]$, let $||g|| = [\int_0^1 g^2(z)dz]^{1/2}$ and $||g||_c = [\int_0^1 (g(z) - \int g(z)dz)^2 dz]^{1/2}$ be the respective L_2 norms of g and its centered version. They measure how much the function g differs from zero or from a constant function, respectively. In this paper, we consider estimation of m_j^0 and $(m_j^0)^{(1)}$, $1 \leq j \leq p$ for sparse high dimensional varying coefficient models where sparsity is defined on a functional level (in L_2 sense). Adapting the idea of the group LASSO to our context, we propose to minimize the following penalized criterion:

$$(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^{(1)}) = \underset{\mathbf{m}, \mathbf{m}^{(1)}}{\operatorname{argmin}}\ L(\mathbf{m}, \mathbf{m}^{(1)}) + \lambda_1 P(\mathbf{m}, \mathbf{m}^{(1)}), \tag{2.2}$$

where $P(\mathbf{m}, \mathbf{m}^{(1)}) = \sum_{j=1}^p \sqrt{\|m_j\|^2 + h^2 \|m_j^{(1)}\|^2}$ for any $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$. Here, $\lambda_1 > 0$ is a regularization parameter. The penalty $\sqrt{\|m_j\|^2 + h^2 \|m_j^{(1)}\|^2}$ jointly controls both, for sparsity of the coefficient function (m_j^0) and for sparsity of its derivative $(m_j^0)^{(1)}$. It contains the rescaled factor h^2 for technical reasons. Our proposal (2.2) is different from the penalized local linear method in Kong et al. [22]. In that paper a penalized criterion for a fixed value of z is considered, see (5) in Kong et al. [22]. This simplifies the asymptotic treatment of the estimator but the chosen set of non-zero coefficient functions depends on the value of z so that it is not applicable for our purpose of estimation under sparsity on the functional level.

It is well known that penalized estimators which employ the LASSO or the group LASSO may fail to achieve consistency in model selection, see also Section 3.2. For this reason, we consider a penalized estimator that corrects for this. Further, the method should be able to discriminate between varying coefficient functions $m_j^0(z)$ over $z \in [0,1]$ with $\|m_j^0\| \neq 0$ and $\|m_j^0\|_c \neq 0$, nonzero constant functions m_j^0 with $\|m_j^0\| \neq 0$ and $\|m_j^0\|_c = 0$ and zero functions $m_j^0 \equiv 0$ with $\|m_j^0\| = 0$ and $\|m_j^0\|_c = 0$. In the first case, we say the coefficients m_j^0 are varying, in the second case, that they are non-varying. We now propose a procedure of estimating the coefficient functions that performs this discrimination. For this purpose we adapt the idea of a group SCAD penalty to our setting. Our version of the SCAD estimator $(\hat{\mathbf{m}}, \hat{\mathbf{m}}^{(1)})$ is defined as the minimizer of the following criterion:

$$L(\mathbf{m}, \mathbf{m}^{(1)}) + \sum_{j=1}^{p} v_j \{\|m_j\|^2 + h^2 \|m_j^{(1)}\|^2\}^{1/2} + \sum_{j=1}^{p} w_j \{\|m_j\|_c^2 + h^2 \|m_j^{(1)}\|^2\}^{1/2}, \tag{2.3}$$

where $v_j = p'_{\lambda_2}((\|\tilde{m}_j\|^2 + h^2 \|\tilde{m}_j^{(1)}\|^2)^{1/2})$, $w_j = p'_{\lambda_2^*}((\|\tilde{m}_j\|_c^2 + h^2 \|\tilde{m}_j^{(1)}\|^2)^{1/2})$ and $p_\lambda(\cdot)$ is the derivative of the SCAD penalty function with regularization parameter $\lambda > 0$, which is given by

$$p'_\lambda(x) = \lambda \left\{ I(x \leq \lambda) + \frac{(\gamma\lambda - x)_+}{(\gamma - 1)\lambda} I(x > \lambda) \right\} \tag{2.4}$$

for some $\gamma > 2$ and $x > 0$. In our simulations and in our data example, $\gamma = 3.7$ is chosen according to the suggestion of Fan and Li [12]. Instead of the SCAD itself, we use a linear approximation of the SCAD penalty (around suitable initial estimates, e.g. the minimizer of (2.2)) in order to overcome difficulties due to non-convexity of the SCAD penalty [46].

Remark 2.1. *In this paper, we consider three possibilities for coefficients: varying, non-varying and zero functions, respectively. Note that in the latter two cases, the derivative of the coefficient function is equal to zero. The penalties in (2.2) and (2.3) include the L_1 norm of the derivative of the estimate. This makes the estimated function exactly equal to a constant function if the coefficient function does not differ too much from a constant function.*

The Bayesian information criterion (BIC) has been used for consistent model selection in linear models. In recent years, it has been proposed as a method of selecting regularization parameters for penalized methods. Work on linear models with high-dimensional settings include Wang and Leng [34], Wang et al. [36, 37], Lee et al. [23]. For our semiparametric setting, we propose the following version of BIC for criterion (2.3): first, we only consider choices $\lambda_2 = \lambda_2^*$. This is done to get a stable choice of the regularization parameter. The value of $\lambda_2 = \lambda_2^*$ is chosen which minimizes

$$\mathrm{BIC}(\lambda_2) = \log L(\hat{\mathbf{m}}_{\lambda_2}, \hat{\mathbf{m}}_{\lambda_2}^{(1)}) + C_n [df_{\lambda_2, V} \frac{\log(nh)}{nh} + df_{\lambda_2, I} \frac{\log n}{n}]. \tag{2.5}$$

Here, the estimators $\hat{\mathbf{m}}_{\lambda_2} = (\hat{m}_{\lambda_2,1}, \ldots, \hat{m}_{\lambda_2,p})^\top$ and $\hat{\mathbf{m}}_{\lambda_2}^{(1)} = (\hat{m}_{\lambda_2,1}^{(1)}, \ldots, \hat{m}_{\lambda_2,p}^{(1)})^\top$ are defined as the minimizer of (2.3) with $\lambda_2 = \lambda_2^*$. Furthermore, $C_n > 0$ is a

sequence of positive constants whose choice will be discussed below. The terms $df_{\lambda_2,V}$ and $df_{\lambda_2,I}$ are the estimated numbers of varying and non-varying coefficients, respectively. That is, $df_{\lambda_2,V} = |\hat{V}_{\lambda_2}|$ and $df_{\lambda_2,I} = |\hat{A}_{\lambda_2} \setminus \hat{V}_{\lambda_2}|$ with estimated index sets $\hat{A}_{\lambda_2} = \{j = 1, \ldots, p : \|\hat{m}_{\lambda_2,j}\| \neq 0\}$ and $\hat{V}_{\lambda_2} = \{j \in \hat{A}_{\lambda_2} : \|\hat{m}_{\lambda_2,j}\|_c \neq 0\}$ of nonzero and varying coefficient functions, respectively. When calculating the BIC in (2.5), the effective sample size nh is used for the number of the nonparametric components instead of the original sample size n. For fixed p the BIC with $C_n = 1$ guarantees consistent model selection, but it may fail to work when p increases, see also Chen and Chen [4] and Lee et al. [23]. Using the ideas developed in Wang et al. [37] and Lee et al. [23] for high dimensional linear models, we consider a diverging constant $C_n \to \infty$ as $n \to \infty$ for high dimensional cases. We will see that a proper choice of C_n leads to consistency of the proposed BIC even in high dimension. See the assumption (A11) and the discussions in Section 3.4.

Although we propose the BIC in (2.5) primarily for selecting λ_2 in our penalization (2.3), the idea also applies to a direct selection problem of the index sets V and I of varying and non-varying coefficient functions. This is done by minimizing the following BIC:

$$\mathrm{BIC}(V, I) = \log L(\bar{\mathbf{m}}_{V,I}, \bar{\mathbf{m}}^{(1)}_{V,I}) + C_n \left(|V| \frac{\log(nh)}{nh} + |I| \frac{\log(n)}{n} \right). \tag{2.6}$$

Here the minimum runs over subsets V and I of $\{1, \ldots, p\}$ with $V \cap I = \emptyset$, and $(\bar{\mathbf{m}}_{V,I}, \bar{\mathbf{m}}^{(1)}_{V,I})$ are estimators of the coefficient functions and their derivatives with index sets of varying and invaring coefficients equal to (V, I). Equivalently, we take $(\bar{\mathbf{m}}_{V,I}, \bar{\mathbf{m}}^{(1)}_{V,I})$ as the minimizer of $L(\mathbf{m}, \mathbf{m}^{(1)})$ subject to the constraints that $\|m_j\| = \|m^{(1)}_j\| = 0$ for $j \in (V \cup I)^c$ and $\|m_j\|_c = \|m^{(1)}_j\| = 0$ for $j \in I$. A similar type of estimator has been discussed in Xia et al. [39] in an i.i.d. setting with fixed dimension. As far as we know, a statistical procedure for simultaneous identification has never been established even in fixed dimensional cases. Only statistical methods for discriminating between zero functions and varying functions, and methods for identifying non-varying coefficients among nonzero functions have been developed [see 39, 35, 19, 44, for example]. Because calculation of $\mathrm{BIC}(V, I)$ over all sets (V, I) is too complex we propose to let $\mathrm{BIC}(V, I)$ only run over sets (V, I) chosen by $\hat{\mathbf{m}}_{\lambda_2}, \hat{\mathbf{m}}^{(1)}_{\lambda_2}$ with λ_2 in an appropriate set of values. We have checked the performance of this estimator in our simulation study. We will show consistency of the proposed BICs in (2.5) and (2.6) in S ection 3.4.

3. Theoretical properties

3.1. Oracle inequality

Let $A^0 = \{1 \leq j \leq p : \|m^0_j\| \neq 0\}$ be the true active set with cardinality $a^0 \equiv |A^0|$. For any p-tuples of (square integrable) functions $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and

$\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$, we define $P(\mathbf{m}, \mathbf{m}^{(1)}) = \sum_{j=1}^p [\|m_j\|^2 + h^2\|m_j^{(1)}\|^2]^{1/2}$, $P_{A^0}(\mathbf{m}, \mathbf{m}^{(1)}) = \sum_{j\in A^0} [\|m_j\|^2 + h^2\|m_j^{(1)}\|^2]^{1/2}$ and $P_{(A^0)^c}(\mathbf{m}, \mathbf{m}^{(1)}) = \sum_{j\notin A^0} [\|m_j\|^2 + h^2\|m_j^{(1)}\|^2]^{1/2}$. For our theoretical analysis we will make use of the following assumptions.

(A1) It holds that $\max_{i,j} |X_i^{(j)}| \le d_n$, a.s.

(A2) The kernel K is a symmetric probability density function with support $[-1, 1]$ and it is Lipschitz continuous.

(A3) There exists a constant $C_1 > 0$, not depending on n, such that

$$\sup_{j\in A^0} |m_j^0(z) - m_j^0(z') - (m_j^0)^{(1)}(z')(z - z')| \le C_1|z - z'|^2$$

for all $z, z' \in [0, 1]$.

(A4) There exists a constant $\phi_n > 0$ such that with probability tending to one

$$\phi_n^2 \left(\sum_{j\in A^0} \|m_j\|^2 + h^2\|m_j^{(1)}\|^2 \right) \le n^{-1} \int \sum_{i=1}^n \left[\mathbf{X}_i^\top \big(\mathbf{m}(z) + \mathbf{m}^{(1)}(z)(Z_i - z)\big) \right]^2 K_h(z - Z_i)dz \quad (3.1)$$

for any $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$ satisfying $P_{(A^0)^c}(\mathbf{m}, \mathbf{m}^{(1)}) \le 3P_{A^0}(\mathbf{m}, \mathbf{m}^{(1)})$.

Assumption (A1) can be relaxed. We only need that $\max_{i,j} |X_i^{(j)}| \le d_n$ with probability tending to 1 as $n \to \infty$. The order of magnitude of d_n in (A1) depends on the tail probability of $X_i^{(j)}$. For example, the maximum of sub-exponential random variables $X_i^{(j)}$ is uniformly bounded by a log factor. In many applications, the $X_i^{(j)}$'s are bounded by a constant, that is, $d_n \le const$ for some constant $const < \infty$. Assumption (A2) is standard. Assumption (A3) holds if the true coefficient functions m_j^0 are twice differentiable on $[0, 1]$ and their 2nd derivatives $(m_j^0)^{(2)}$ are uniformly bounded, i.e $\sup_{j\in A^0} \sup_{z\in[0,1]} |(m_j^0)^{(2)}(z)| < \infty$. When p is fixed, it has typically been assumed that there exists a constant $\phi > 0$ such that

$$\begin{aligned} S(\mathbf{m}, \mathbf{m}^{(1)}) &= n^{-1} \int \sum_{i=1}^n \left[\mathbf{X}_i^\top \big(\mathbf{m}(z) + \mathbf{m}^{(1)}(z)(Z_i - z)\big) \right]^2 K_h(z - Z_i)dz \\ &\ge \phi^2 \left(\sum_{j=1}^p \|m_j\|^2 + h^2\|m_j^{(1)}\|^2 \right) \end{aligned} \quad (3.2)$$

for $\mathbf{m}(\cdot) = (m_1(\cdot), \ldots, m_p(\cdot))$ and $\mathbf{m}(\cdot)^{(1)} = (m_1(\cdot)^{(1)}, \ldots, m_p(\cdot)^{(1)})$ (with large probability). However, for very large p it may be too restrictive to assume (3.2) for all $(\mathbf{m}, \mathbf{m}^{(1)})$. For (A4), we adapt the concept of 'compatibility condition'

that has been developed for high dimensional models [see 1, for example]. For a general comparison of different conditions on design matrices in high dimensional linear models see van de Geer and Bühlmann [30]. There it has also been pointed out that their version of the 'compatibility condition' allows for a fairly general class of design matrices. Since assumption (A4) depends on the data, we will discuss a population version of (A4) later.

The following lemma and theorem state oracle results for the estimator $(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^{(1)})$ defined at (2.2). Define $\mathcal{T} = \mathcal{T}_1 \cap \mathcal{T}_2 \cap \mathcal{T}_3$ where

$$\begin{aligned}\mathcal{T}_1 \equiv \mathcal{T}_1(\lambda_0) = \Bigg\{ & \left| n^{-1} \sum_{i=1}^{n} \int \epsilon_i \mathbf{X}_i^\top [\mathbf{m}(z) + \mathbf{m}^{(1)}(z)(Z_i - z)] K_h(z - Z_i) dz \right| \\ & \leq \lambda_0 \sum_{j=1}^{p} \left(\|m_j\| + h \|m_j^{(1)}\| \right) \\ & \text{for all } \mathbf{m} = (m_1, \ldots, m_p)^\top \text{ and } \mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top \Bigg\}, \\ \mathcal{T}_2 \equiv \mathcal{T}_2(C_2) = & \left\{ \sup_{z \in [0,1]} \sup_{1 \leq j,k \leq p} n^{-1} \sum_{i=1}^{n} \left| X_i^{(j)} X_i^{(k)} K_h(z - Z_i) \right| \leq C_2 \right\}.\end{aligned}$$

and where $\mathcal{T}_3$ is the event that (3.1) holds for $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$ satisfying $P_{(A^0)^c}(\mathbf{m}, \mathbf{m}^{(1)}) \leq 3 P_{A^0}(\mathbf{m}, \mathbf{m}^{(1)})$.

Lemma 3.1. *Under (A1) and (A2), there exists a constant $M > 0$ such that*

$$P(\mathcal{T}_1) \geq 1 - M \lambda_0^{-2} \frac{\log p}{nh} d_n^2.$$

Theorem 3.1. *Suppose that the assumptions (A1)–(A4) hold and that $\lambda_1 \geq 4\sqrt{2}(\lambda_0 + C_1 C_2 a^0 h^2)$. Then, on $\mathcal{T}$,*

$$S\big(\tilde{\mathbf{m}} - \mathbf{m}^0, \tilde{\mathbf{m}}^{(1)} - (\mathbf{m}^0)^{(1)}\big) + \lambda_1 P\big(\tilde{\mathbf{m}} - \mathbf{m}^0, \tilde{\mathbf{m}}^{(1)} - (\mathbf{m}^0)^{(1)}\big) \leq 4 \lambda_1^2 a^0 / \phi_n^2.$$

Below we will state assumptions under which $P(\mathcal{T}_2) \to 1$, see Theorem 3.2. Here and below, we write $a_n \approx b_n$ for two sequences a_n and b_n if the ratio a_n / b_n is bounded away from zero and infinity. Suppose $nh \to \infty$ and $h \to 0$ as $n \to \infty$. Taking $\lambda_0 \approx d_n \sqrt{\log p / (nh)}$ from Lemma 3.1, Theorem 3.1 gives

$$\begin{aligned}\sum_{j=1}^{p} \|\tilde{m}_j - m_j^0\| &\sim \left(d_n \sqrt{\frac{\log p}{nh}} + a^0 h^2 \right) a^0 \phi_n^{-2}, \\ \sum_{j=1}^{p} \|\tilde{m}_j^{(1)} - (m_j^0)^{(1)}\| &\sim \left(d_n \sqrt{\frac{\log p}{nh^3}} + a^0 h \right) a^0 \phi_n^{-2}.\end{aligned}$$

Moreover, suppose that $X_i^{(j)}$ are bounded by a log-factor, that ϕ_n is bounded away from zero and that the cardinality of the true active set A^0 is of order

$(\log p)^\gamma$ for some $\gamma > 0$. Then, up to a log term, the above rates coincide with that of oracle estimators that utilize knowledge of the set A^0, and that achieve the optimal nonparametric convergence rate when $h \approx n^{-1/5}$. Thus, our results can be interpreted as oracle results for the estimators of the coefficient functions and their derivatives. Regarding model selection, we will show in Section 3.2 that the LASSO estimator $(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^{(1)})$ (with any choice of λ_1) cannot achieve consistency in general. Thus, we chose λ_1 which minimizes an estimate of prediction error in the simulated and real data examples.

We present theoretical results for varying coefficient models with i.i.d. data and for time varying regression models. We introduce some generic notations where the definitions differ in these two settings. We define $\mathbf{\Sigma}(z) = E[\mathbf{X}\mathbf{X}^\top | Z = z]$ under the i.i.d. setting and $\mathbf{\Sigma}(i/n) = E[\mathbf{X}_i\mathbf{X}_i^\top]$ under the time series setting. Furthermore, f denotes the density of Z under the i.i.d. setting and $f(z) \equiv 1$ under the time series setting. We now state sufficient conditions for (A4). Note that the quantity (3.1) in the assumption (A4) depends on the data $\{(\mathbf{X}_i, Z_i, Y_i), 1 \le i \le n\}$. We now state an assumption that is related to (A4), but with random quantities replaced by nonrandom terms:

(A4') There exists $\phi_n' > 0$ such that

$$\phi_n'^2 \left(\sum_{j \in A^0} \|m_j\|^2 + h^2 \|m_j^{(1)}\|^2 \right) \le \int \mathbf{m}(z)^\top \mathbf{\Sigma}(z) \mathbf{m}(z) f(z) dz + h^2 \int \mathbf{m}^{(1)}(z)^\top \mathbf{\Sigma}(z) \mathbf{m}^{(1)}(z) f(z) dz \quad (3.3)$$

for any $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$ satisfying $P_{(A^0)^c}(\mathbf{m}, \mathbf{m}^{(1)}) \le 3 P_{A^0}(\mathbf{m}, \mathbf{m}^{(1)})$.

In our notation, $0 < C < \infty$ denotes a generic constant, not depending on n. This means that the variable name C is used for different constants, even in the same equation. We now state additional assumptions. We will show below that under (A1)–(A2) and under these conditions, (A4') implies (A4).

(A5) There exists a constant $0 < C < \infty$, not depending on n, such that

$$\sup_{1 \le j,k \le p} |\Sigma_{jk}(z) - \Sigma_{jk}(z'))| \le C|z - z'|$$

for any $z, z' \in [0,1]$ and $\sup_{1 \le j \le p} \sup_z \Sigma_{jj}(z) \le C < \infty$.

(A6) Under the i.i.d. setting, $n^{-1}h^{-2} \log n \to 0$ and the density f of Z is Lipschitz continuous and bounded away from zero; under the time series setting, $\{\mathbf{X}_i, 1 \le i \le n, n \in \mathbb{Z}\}$ is α-mixing where the mixing coefficients

$$\alpha(k) = \sup_{(m,n): m \le n-k} \sup_{\substack{A \in \sigma(\mathbf{X}_s, s \le m) \\ B \in \sigma(\mathbf{X}_s, s \ge m+k)}} |P(A \cap B) - P(A)P(B)|$$

satisfy $\alpha(k) \le Ck^{-\alpha}$ for some $C > 0$ and $\alpha > 1$, and

$$np^2 d_n^{-2} \left(\frac{\log n + \log p}{nh} \right)^{\alpha/2} \to 0, \quad n^{-1}h^{-2} \to 0. \quad (3.4)$$

(A7) Under the i.i.d. setting, $(\phi_n')^{-2}a^0\left(d_n^2(\log n+\log p)^{1/2}(nh)^{-1/2}+h\right)\to 0$; under the time series setting, $(\phi_n')^{-2}a^0\Big(d_n^2(\log n+\log p)^{1/2}(nh)^{-1/2}+h+n^{-1}h^{-2}\Big)\to 0$.

Since the constants ϕ_n and ϕ_n' in the assumptions are not unique, we suppose that ϕ_n and ϕ_n' are chosen as the largest positive constants satisfying (A4) and (A4'), respectively. In (A6), the first condition $n^{-1}h^{-2}\log n\to 0$ for the i.i.d. settings guarantees uniform bounds on $N(z)$ of the order nh, with probability tending to one, where $N(z)$ is the number of Z_i's that fall into the interval $[z-h,z+h]$, see the discussion in the first paragraph of the Appendix for details. Note that since $Z_i=i/n$ under the time series setting, $N(z)\le 2nh+1$ for $z\in[0,1]$. Under the time series setting, the mixing condition of (A6) is not strong, compare also recent work on local stationary processes [see 17, 32, for example]. Time dependency of the covariates $\mathbf{X}$ in the time varying coefficient model restricts the growth rate of p. The first condition in (3.4) implies that the order of p does not exceed $(nh)^{\alpha/4}$ and thus, the larger (smaller) α is, the more (less) covariates are allowed in the model. If the α-mixing coefficients $\alpha(k)$ decrease exponentially and p grows at any polynomial rate of n, i.e. $p=O(n^\kappa)$ for $\kappa>0$ then there is no such restriction on p as long as (A7) holds. Then, note that because of $\alpha(k)\le Ck^{-\alpha}$ for all $\alpha>0$ the first condition in (3.4) is automatically satisfied. The assumption (A7) implies that the number a^0 of true nonzero coefficient functions cannot grow too fast. In the i.i.d. setting, it allows for ultra-high dimensionality of the variables, i.e., $p=o(\exp(nh))$ if ϕ_n, a^0, d_n are bounded.

The following theorem states an asymptotic equivalence between ϕ_n and ϕ_n'.

Theorem 3.2. *Assume that (A1)–(A2), (A4') and (A5)–(A7) hold. Then, (A4) holds with a sequence ϕ_n that fulfills $C\phi_n'\le\phi_n\le C^{-1}\phi_n'$ for some $C>0$.*

Let $b_n=\left(d_n(\log p)^{1/2}(nh)^{-1/2}+a^0h^2\right)a^0(\phi_n')^{-2}$. Theorem 3.2 also gives a uniform rate of convergence for the estimators $\tilde m_j$ and $\tilde m_j^{(1)}$, $j=1,\dots,p$.

Corollary 3.1. *Under (A1)–(A3),(A4'), (A5)–(A7) and $b_n\to 0$, we have that* $\sum_{j=1}^p\|\tilde m_j-m_j^0\|+h\|\tilde m_j^{(1)}-(m_j^0)^{(1)}\|=O_p(b_n)$.

3.2. Consistency and inconsistency of group LASSO estimators

In this section, we study if the proposed estimator $(\tilde{\mathbf{m}},\tilde{\mathbf{m}}^{(1)})$ in (2.2) achieves consistency in model selection. Here, consistency means that the selected set $\tilde A=\{j=1,\dots,p:\|\tilde m_j\|\neq 0\}$ by the estimator is equal to the true active set A^0 with probability tending to 1 as $n\to\infty$. Discrimination between varying and non-varying functions in model (2.1) would require an additional model choice procedure. In this section we will state a condition that is necessary for consistency, see Proposition 3.1 and Theorem 3.3. At the end of this section, we will use these results to show inconsistency of our group LASSO $(\tilde{\mathbf{m}},\tilde{\mathbf{m}}^{(1)})$ in an example.

For simplification we make the following additional condition:

(C) the cardinality $a^0 = |A^0|$ of the true active set is fixed and the smallest eigenvalues of $\mathbf{\Sigma}_{A^0,A^0}(z)$ are bounded away from zero uniformly in $z \in [0,1]$, that is, there exists a constant $\phi^2 > 0$, not depending on z, such that $\mathbf{a}^\top \mathbf{\Sigma}_{A^0,A^0}(z)\mathbf{a} \geq \phi^2 \mathbf{a}^\top \mathbf{a}$ for any $\mathbf{a} \in \mathbb{R}^{|A^0|}$.

Before presenting our theoretical results, we introduce some notation. Let $\mathbf{\Gamma}_i(z) = (\mathbf{X}_i^\top, ((Z_i - z)/h)\mathbf{X}_i^\top)^\top$ and $\hat{\mathbf{S}}(z) = n^{-1}\sum_{i=1}^n \mathbf{\Gamma}_i \mathbf{\Gamma}_i^\top K_h(z - Z_i)$. We also skip the argument and write $\mathbf{\Gamma}_i$ and $\hat{\mathbf{S}}$. Define $\mathbf{\Gamma}_{i,j}$ and $\mathbf{\Gamma}_{i,A^0}$ as $\mathbf{\Gamma}_i$ but with $\mathbf{X}_i$ replaced by $X_i^{(j)}$ and $\mathbf{X}_{i,A^0} = (X_i^{(j)} : j \in A^0)^\top$, respectively. Similarly, $\hat{\mathbf{S}}_{j,A^0} = n^{-1}\sum_{i=1}^n \mathbf{\Gamma}_{i,j}\mathbf{\Gamma}_{i,A^0}^\top K_h(\cdot - Z_i)$ and $\hat{\mathbf{S}}_{A^0,A^0} = n^{-1}\sum_{i=1}^n \mathbf{\Gamma}_{i,A^0}\mathbf{\Gamma}_{i,A^0}^\top K_h(\cdot - Z_i)$. Given any $0 \leq z \leq 1$ and any $j = 1, \ldots, p$, we define $\delta^{(j)}(z) = n^{-1}\sum_{i=1}^n [\mathbf{\Gamma}_{i,j}(z) - \hat{\mathbf{S}}_{j,A^0}(z)\hat{\mathbf{S}}_{A^0,A^0}^{-1}(z)\mathbf{\Gamma}_{i,A^0}(z)]e_i^0(z)K_h(z - Z_i)$ also denoted by $(\delta_{1,j}, \delta_{2,j})^\top$, where $e_i^0(z) = Y_i - \mathbf{X}_i^\top[\mathbf{m}^0(z) + (\mathbf{m}^0)^{(1)}(z)(Z_i - z)]$. Let $\tilde{\mathbf{s}}_{A^0}(\cdot) = (\tilde{s}_j(\cdot) : j \in A^0)^\top$ and $\tilde{\mathbf{s}}_{A^0}^{(1)}(z) = (\tilde{s}_j^{(1)}(z) : j \in A^0)^\top$ with $\tilde{s}_j(z) = \tilde{m}_j(z)/\sqrt{\|\tilde{m}_j\|^2 + h^2\|\tilde{m}_j^{(1)}\|^2}$ and $\tilde{s}_j^{(1)}(z) = h\tilde{m}_j^{(1)}(z)/\sqrt{\|\tilde{m}_j\|^2 + h^2\|\tilde{m}_j^{(1)}\|^2}$.

The following proposition and theorem give a necessary condition for consistency of the group LASSO.

Proposition 3.1. *Suppose that (A1)–(A2), (C) and (A5)–(A7) hold and that* $\lim_{n\to\infty} P(\tilde{A} = A^0) = 1$. *Then, for any* $\varepsilon > 0$, *we have*

$$\sqrt{\int [\mathbf{\Sigma}_{j,A^0}(z)\mathbf{\Sigma}_{A^0,A^0}(z)^{-1}\tilde{\mathbf{s}}_{A^0}(z)]^2 + [\mathbf{\Sigma}_{j,A^0}(z)\mathbf{\Sigma}_{A^0,A^0}(z)^{-1}\tilde{\mathbf{s}}_{A^0}^{(1)}(z))]^2 dz}$$
$$\leq 1 + \varepsilon + 2\lambda_1^{-1}\sqrt{\int \delta_{1,j}^2 + \delta_{2,j}^2} \tag{3.5}$$

for $j \notin A^0$ *with probability tending to 1 as* $n \to \infty$.

Let $\Delta_{j,s}^2 = \int \{n^{-1}\sum_{i=1}^n X_i^{(j)}\epsilon_i((Z_i - z)/h)^s K_h(z - Z_i)\}^2 dz$ for $1 \leq j \leq p$ and $s = 0, 1$. Define $\mathcal{S}_1 \equiv \mathcal{S}_1(\lambda_0) = \{\max_{1\leq j\leq p,\ s=0,1} \Delta_{j,s}^2 \leq \lambda_0^2\}$ and $\mathcal{S}_1' \equiv \mathcal{S}_1'(\lambda_0') = \{\max_{j\in A^0,\ s=0,1} \Delta_{j,s}^2 \leq \lambda_0'^2\}$. Then, $\mathcal{S}_1(\lambda_0) \subset \mathcal{T}_1(\lambda_0)$ and similarly as in the proof of Lemma 3.1, it can be shown that $P(\mathcal{S}_1) \geq 1 - M\lambda_0^{-2}(nh)^{-1}d_n^2 \log p$ for some $M > 0$.

Theorem 3.3. *Suppose that assumptions (A1)–(A3), (C) and (A5)–(A7) hold. Then, the following properties are obtained: (i) on* $\mathcal{S}_1' \cap \mathcal{S}_1 \cap \mathcal{T}_2$ *with* $C_2 > C$, $\max_{1\leq j\leq p}\sqrt{\int \delta_{1,j}^2 + \delta_{2,j}^2} \leq \sqrt{2}(\lambda_0 + \phi^{-2}(a^0 C_2)\lambda_0' + C_1 C_2 a^0 h^2)$; *(ii)* $P(\mathcal{S}_1') \geq 1 - M\lambda_0'^{-2}(nh)^{-1}\ d_n^2$ *for some* $M > 0$ *where* C_2, C *are the constants in the definition of* $\mathcal{T}_2$ *and in assumption (A5), respectively.*

Remark 3.1. *For* $p \to \infty$ *with asymptotically optimal choices of* λ_0 *and* λ_0', $\lambda_0 \approx d_n\sqrt{\log p/(nh)}$ *and* $\lambda_0' \approx d_n\sqrt{1/(nh)}$, *(3.5) is uniformly bounded by* $3/2 +$

2ε *(with probability tending to one) as long as the conditions in Theorem 3.3 are fulfilled and it holds that* $\lambda_1 \geq 4\sqrt{2}(\lambda_0 + C_1 C_2 a^0 h^2)$*, see Theorem 3.1. It easily follows that*

$$\sqrt{\int [\boldsymbol{\Sigma}_{j,A^0}(z)\boldsymbol{\Sigma}_{A^0,A^0}(z)^{-1}\mathbf{s}^0_{A^0}(z)]^2 dz} \leq 3/2, \tag{3.6}$$

where $\mathbf{s}^0_{A^0}(\cdot) = (s^0_j(\cdot) : j \in A^0)^\top$ *and* $s^0_j(\cdot) = m^0_j(\cdot)/\sqrt{\|m^0_j\|^2}$.

Remark 3.2. *For fixed p we get the following result under the conditions of Theorem 3.3 and under the additional assumption that the largest eigenvalues of* $E(\epsilon\epsilon^\top|\mathbf{X}_1, \ldots, \mathbf{X}_n, Z_1, \ldots, Z_n)$ *are bounded by a constant, not depending on n and on the values of* $\mathbf{X}_1, \ldots, \mathbf{X}_n, Z_1, \ldots, Z_n$*. Here* ϵ *denotes the vector* $(\epsilon_1, \ldots, \epsilon_n)^\top$*. Then, on* $\mathcal{S}_1 \cap \mathcal{T}_2$, $\max_{1\leq j\leq p}\sqrt{\int \delta^2_{1,j} + \delta^2_{2,j}} \leq \sqrt{2}(\lambda_0 + C_1 C_2 a_0 h^2)$ *and* $P(\mathcal{S}_1) \geq 1 - M\lambda_0^{-2}(nh)^{-1}\, d_n^2$ *for some* $M > 0$*. This also implies (3.6) for fixed p.*

We now give an example where our group LASSO is inconsistent. Suppose that the matrix $\boldsymbol{\Sigma}(z)$ is constant over z, i.e., $\boldsymbol{\Sigma}(z) \equiv \boldsymbol{\Sigma}$. Let $\mathbf{1}_K = (1, \ldots, 1)^\top \in \mathbb{R}^K$. Similarly as in Corollary 1 in Zou [45], we choose $A^0 = \{1, \ldots, a^0\}$ with $a^0 \geq 2$, $\boldsymbol{\Sigma}_{A^0,A^0} = (1-\rho_1)I + \rho_1 \mathbf{1}_{a^0}\mathbf{1}^\top_{a^0}$, $\boldsymbol{\Sigma}_{a^0+1,A^0} = \rho_2 \mathbf{1}^\top_{a^0}$, $\boldsymbol{\Sigma}_{jj} = 1$, $j \notin A^0$ and $\boldsymbol{\Sigma}_{jk} = \boldsymbol{\Sigma}_{kj} = 0$ for $j \geq a^0 + 2$ and $k \leq a^0 + 1$. In this model there is one irrelevant predictor $X^{(a^0+1)}$ correlated with relevant predictors. It holds that

$$\sqrt{\int [\boldsymbol{\Sigma}_{a^0+1,A^0}(z)\boldsymbol{\Sigma}_{A^0,A^0}(z)^{-1}\mathbf{s}_{A^0}(z)]^2 dz} = \left|\frac{\rho_2}{1 + (a^0-1)\rho_1}\right| \sqrt{\int (\mathbf{1}^\top_{a^0}\mathbf{s}_{A^0}(z))^2 dz}$$

because of $\boldsymbol{\Sigma}_{a^0+1,A^0}\boldsymbol{\Sigma}^{-1}_{A^0,A^0} = \rho_2/(1 + (a^0-1)\rho_1)\mathbf{1}^\top_{a^0}$. Therefore, if $-(a^0-1)^{-1} < \rho_1 < -(a^0)^{-1}$, $1 + (a^0-1)\rho_1 < |\rho_2| < \sqrt{(1 + (a^0-1)\rho_1)/a^0}$ and if the nonvanishing functions m_j, $j \in A^0$ are all nonnegative (or nonpositive), then, the model selection via the method $(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^{(1)})$ defined at (2.2) is not consistent because condition (3.6) does not hold.

3.3. Oracle properties

In this section, we present oracle properties of the estimator $(\hat{\mathbf{m}}, \hat{\mathbf{m}}^{(1)})$, defined as minimizer of (2.3). Put $\hat{\mathbf{m}}_{A^0} = (\hat{m}_j : j \in A^0)^\top$ and $\hat{\mathbf{m}}^{(1)}_{A^0} = (\hat{m}^{(1)}_j : j \in A^0)^\top$. Furthermore, define $V^0 = \{j \in A^0 : \|m^0_j\|_c \neq 0\}$ as the index set of true coefficient functions that are varying over $z \in [0,1]$. We compare $\hat{\mathbf{m}}_{A^0}$ and $\hat{\mathbf{m}}^{(1)}_{A^0}$ with the oracle estimators $\hat{\mathbf{m}}^{ora} = (\hat{m}^{ora}_j : j \in A^0)^\top$ and $(\hat{\mathbf{m}}^{ora})^{(1)} = ((\hat{m}^{ora}_j)^{(1)} : j \in A^0)^\top$ that are defined as minimizers of

$$n^{-1}\int \sum_{i=1}^n [Y_i - \sum_{j\in A^0} X_i^{(j)}(m_j(z) + m_j^{(1)}(z)(Z_i - z))]^2 K_h(z - Z_i)\, dz \tag{3.7}$$

with respect to $m_j, (m_j)^{(1)}$ for $j \in A^0$ under the constraint that m_j are constant functions for $j \in A^0 \setminus V^0$ and that $(m_j)^{(1)} \equiv 0$ for $j \in A^0 \setminus V^0$. The oracle estimator is an infeasible estimator that makes use of the unknown true index sets A^0 and V^0.

For the asymptotic analysis in this section we make the following additional assumption.

(A8) There exists a positive constant $\delta > 0$ such that $\inf_{j \in A^0} \|m_j^0\| > \delta$ and $\inf_{j \in V^0} \|m_j^0\|_c > \delta$.

Theorem 3.4. *Suppose that assumptions (A1)–(A3),(A4′), (A5)–(A8) hold, that* $\max\{\lambda_2, \lambda_2^*\}/\delta \to 0$ *and that* $\min\{\lambda_2, \lambda_2^*\}/b_n \to \infty$. *Then,* (i) $P(\|\hat{m}_j\| = \|\hat{m}_j^{(1)}\| = 0 \text{ for } j \notin A^0 \text{ and } \|\hat{m}_j\|_c = \|\hat{m}_j^{(1)}\| = 0 \text{ for } j \notin V^0) \to 1$ *as* $n \to \infty$*;* (ii) *with probability tending to 1 as* $n \to \infty$, $(\hat{\mathbf{m}}_{A^0}, \hat{\mathbf{m}}_{A^0}^{(1)})$ *minimizes (3.7) with respect to* $m_j, (m_j)^{(1)}$, $j \in A^0$ *subject to* $\|m_j\|_c = 0$ *and* $\|(m_j)^{(1)}\| = 0$ *for* $j \in A^0 \setminus V^0$.

Theorem 3.4 states that our proposed procedure consistently identifies the true index sets of varying and non-varying coefficients in the model. Thus, the resulting estimators $\hat{m}_j$ and $\hat{m}_j^{(1)}$ of the nonzero coefficient functions have the same asymptotic properties as the oracle estimators $\hat{m}_j^{ora}$ and $(\hat{m}_j^{ora})^{(1)}$ for $j \in A^0$. Using standard arguments of kernel smoothing it can be shown that the estimators $\hat{m}_j$, $j \in A^0 \setminus V^0$ of the (nonzero) constant coefficients achieve the parametric $\sqrt{n}$-rate of convergence under certain regularity conditions, see (A.26) in Appendix A.7. The required assumptions allow $h \approx n^{-1/5}$ for the case that $a^0 = |A^0|$ is fixed. This implies that in this case the same bandwidth $h \approx n^{-1/5}$ can be used to achieve an optimal rate of convergence for both, the parametric and the nonparametric components, at the same time. In contrast to other methods in semiparametrics, undersmoothing is not required for $\sqrt{n}$ consistency of the parametric estimators.

3.4. Consistent identification of BIC

In this section, we study consistency of the BIC methods proposed in (2.5) and (2.6). Let $I^0 = A^0 \setminus V^0$ be the index set of the true non-varying coefficient functions m_j^0 with $\|m_j^0\| \neq 0$ and $\|m_j^0\|_c = 0$ in model (2.1). Given a pair of subsets V and I of $\{1, \ldots, p\}$ with $V \cap I = \emptyset$, define $A \equiv A(V, I) = V \cup I$. For a technical reason, we let BIC run over index sets V of varying coefficient functions and index sets I of non-varying coefficient functions with $|A(V, I)| \leq s_n$, where s_n is chosen such that (A4") and (A10) hold. For similar dimension restrictions in model selection for (high dimensional) linear models compare Chen and Chen [4], Kim et al. [20] and Lee et al. [23]. We put $\mathcal{M} = \{(V, I) : V, I \subset \{1, \ldots, p\},\ V \cap I = \emptyset,\ |A(V, I)| \leq s_n\}$. Note that it is computationaly infeasible to calculate $\text{BIC}(V, I)$ for all elements in $\mathcal{M}$, at least when p is large. A modification where the minimization does not run over the full space $\mathcal{M}$ will be discussed below and studied in the simulation section 4.3. The now developed

theory for the estimator with (V, I) running over the full space $\mathcal{M}$ will be applied to this modification below.

For stating our results on the BIC methods we need the following additional assumptions.

(A4") There exists a constant $\phi'' > 0$, not depending on n, such that with probability tending to one,

$$(\phi'')^2 \left(\sum_{j \in A} \|m_j\|^2 + h^2 \|m_j^{(1)}\|^2 \right) \leq S\big(\mathbf{m}, \mathbf{m}^{(1)}\big)$$

for all $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$ with sets $A = \{1 \leq j \leq p : m_j \not\equiv 0\}$ whose cardinality is bounded by $2s_n$.

(A9) The sequences d_n in (A1) and $a^0 = |A^0|$ are bounded. The constant δ in (A8) is bounded away from zero.

(A10) It holds that $p = O(n^\kappa)$ for some $\kappa > 0$, $h \approx n^{-1/5}$, $a^0 \leq s_n$ and $s_n \log n/(nh) \to 0$.

(A11) The constant C_n in (2.5) and (2.6) satisfies $C_n \to \infty$ and $C_n \log n (nh)^{-1} \to 0$.

(A12) It holds that

$$\max_{1 \leq k \leq p} |n^{-1} \sum_{i=1}^n \int \hat{e}_i^{ora}(z) K_h(z - Z_i) dz X_i^{(k)}| = o_p(\sqrt{C_n \log n/n}),$$

where $\hat{e}_i^{ora}(z) = Y_i - \sum_{j \in A^0} X_i^{(j)} [\hat{m}_j^{ora}(z) + (\hat{m}_j^{ora})^{(1)}(z)(Z_i - z)]$ for $1 \leq i \leq n$ and $z \in [0, 1]$.

For simplicity, we make assumptions (A4") and (A9) that put stronger conditions on ϕ'', d_n, a^0, δ than the assumptions in Subsections 3.1 and 3.3. Our theory can be generalized to cases where the constants ϕ'', d_n, a^0 and δ depend on the sample size n, i.e, ϕ'', δ tend to zero as $n \to \infty$ or d_n, a^0 diverge with n. However, then more restrictive conditions on C_n are needed that depend on the unknown quantities ϕ'', d_n, a^0 and δ. This restricts the practical use of such a result. The (A4") is a modification of the so-called 'sparse Riesz condition' of Zhang and Huang [42] in high dimensional linear regression. For an application of this assumption see also Wang et al. [37] and Lee et al. [23]. We make the assumption (A12) in order to show that our procedures correctly classify the estimated constant coefficients into zeros and non-varying ones. It is also needed for getting asymptotic properties for the oracle estimator $\hat{m}_j^{ora}, (\hat{m}_j^{ora})^{(1)}$ for $j \in A^0$. The asymptotics of the oracle estimator is well understood and the derivation of sufficient high-level conditions follows standard lines, see also Appendix A.7 for a related discussion.

The following theorem states that the BIC method defined in (2.6) consistently estimates the index sets V^0 and I^0.

Theorem 3.5. *Assume that (A1)–(A3), (A4"), (A8)–(A12) hold. Then, we have*

$$P\left(\min_{(V,I)\in\mathcal{M}:V\neq V^0 \text{ or } I\neq I^0} \mathrm{BIC}(V,I) > \mathrm{BIC}(V^0,I^0)\right) \to 1 \quad \text{as} \quad n\to\infty.$$

Remark 3.3. *For the theoretical statement of the theorem, any positive sequence $\{C_n\}$ that fulfills (A11) works. In all our numerical experiments, we used $C_n = \sqrt{\log p}$ and this choice lead to good results.*

Theorem 3.5 can be translated to a consistency result of the BIC in (2.5), which is a penalization parameter selector for the problem (2.3). Define $\hat{\lambda}_2 = \arg\min_{\lambda_2} \mathrm{BIC}(\lambda_2)$ where the 'argmin' runs over all $\lambda_2 > 0$ such the cardinality $|\hat{A}_{\lambda_2}|$ of $\hat{A}_{\lambda_2}$ is smaller than or equal to s_n. Here $\hat{A}_{\lambda_2} = \hat{V}_{\lambda_2} \cup \hat{I}_{\lambda_2}$ is the subset selected by the penalized estimator $\hat{m}_{\lambda_2,j}$ and $\hat{m}^{(1)}_{\lambda_2,j}$, for $1 \le j \le p$. Recall that $\hat{V}_{\lambda_2} = \{j = 1,\dots,p : \|\hat{m}_{\lambda_2,j}\| \neq 0, \|\hat{m}_{\lambda_2,j}\|_c \neq 0\}$ and $\hat{I}_{\lambda_2} = \{j = 1,\dots,p : \|\hat{m}_{\lambda_2,j}\| \neq 0, \|\hat{m}_{\lambda_2,j}\|_c = 0\}$. From Theorem 3.4 and 3.5, one gets that $P(\hat{V}_{\hat{\lambda}_2} = V^0,\ \hat{I}_{\hat{\lambda}_2} = I^0) \to 1$ as $n \to \infty$. That is, the BIC in (2.5) chooses the penalty constant λ_2 such that the resulting estimator $(\hat{\mathbf{m}}_{\lambda_2}, \hat{\mathbf{m}}^{(1)}_{\lambda_2})$ consistently selects the true V^0 and I^0. Furthermore, we get that minimizing $\mathrm{BIC}(V,I)$ over $(\hat{V}_{\lambda_2}, \hat{I}_{\lambda_2})$ leads to a consistent estimator of (V^0, I^0). We denote the minimizing sets by $(\widetilde{V}, \widetilde{I})$. We call the estimator $(\bar{\mathbf{m}}_{\widetilde{V},\widetilde{I}}, \bar{\mathbf{m}}^{(1)}_{\widetilde{V},\widetilde{I}})$ the BIC-estimator. Recall that $(\bar{\mathbf{m}}_{V,I}, \bar{\mathbf{m}}^{(1)}_{V,I})$ is the minimizer of the unpenalized criterion $L(\mathbf{m}, \mathbf{m}^{(1)})$ subject to the constraints that $\|m_j\| = \|m_j^{(1)}\| = 0$ for $j \in (V \cup I)^c$ and $\|m_j\|_c = \|m_j^{(1)}\| = 0$ for $j \in I$. Thus we have three types of estimators: the LASSO-estimator, the SCAD-estimator with penalty constant chosen by BIC and the just introduced BIC-estimator. We will compare the three estimators in our simulation study in the next section.

4. Numerical studies

4.1. Numerical implementation

Our proposed criteria (2.2) and (2.3) include integrals over the interval $[0,1]$. In the numerical implementation of the method we propose to approximate the integrals by discretization schemes. In our computations we take J discretization points of the interval $[0,1]$ with $J = 100$ and compute the Riemann sum of the integral for numerical integration. Then our problems turn into a $2Jp$ dimensional optimization problem. The discretized problem of minimizing (2.2) can be formulated as a typical problem of the group LASSO and easily solved by any numerical algorithm for the group LASSO. In contrast, the resulting problem of (2.3) is quite complicated because there is an hierarchical structure between the different penalties $\{\|m_j\|^2 + h^2\|m_j^{(1)}\|^2\}^{1/2}$ and $\{\|m_j\|_c^2 + h^2\|m_j^{(1)}\|^2\}^{1/2}$, more precisely, $\{\|m_j\|_c^2 + h^2\|m_j^{(1)}\|^2\}^{1/2} \le \{\|m_j\|^2 + h^2\|m_j^{(1)}\|^2\}^{1/2}$. Here, we present

a numerical algorithm for minimizing (the discretized version of) the criterion (2.3). The optimization problem (2.2) can be done either by using any available software for solving a group LASSO problem or by applying our algorithm for (2.3) with $v_1 = \cdots = v_p = \lambda$ and $w_1 = \cdots = w_p = 0$.

Recall the definitions of $\mathbf{\Gamma}_i(z)$ and $\hat{\mathbf{S}}(z)$ in Section 3.2, and define $\mathbf{L}(z) = n^{-1}\sum_{i=1}^n \mathbf{\Gamma}_i Y_i K_h(z - Z_i)$. With J discretization points $z_1, \ldots, z_J$ of the interval $[0,1]$, our problem (2.3) can be rewritten as follows:

$$\min_{\mathbf{x},\mathbf{q},\mathbf{s},\mathbf{t}} \left\{ J^{-1}\sum_{j=1}^{J}(q_j - 2\mathbf{L}(z_j)^\top \mathbf{x}(z_j)) + \sum_{k=1}^{p}(v_k s_k + w_k t_k) \right\} \quad (4.1)$$

$$\text{such that} \qquad \begin{aligned} \mathbf{x}(z_j)^\top \hat{\mathbf{S}}(z_j)\mathbf{x}(z_j) &\le q_j, \ j = 1, \ldots J, \\ \|\mathbf{A}_k \mathbf{x}\|_2 &\le s_k, \quad k = 1, \ldots, p, \\ \|\mathbf{B}_k \mathbf{x}\|_2 &\le t_k, \quad k = 1, \ldots, p, \end{aligned}$$

where $\mathbf{x} = (\mathbf{x}(z_1)^\top, \ldots, \mathbf{x}(z_J)^\top)^\top$, $\mathbf{q} = (q_1, \ldots, q_J)^\top$, $\mathbf{s} = (s_1, \ldots, s_p)^\top$ and $\mathbf{t} = (t_1, \ldots, t_p)^\top$. Also $\mathbf{A}_k$ and $\mathbf{B}_k$ denote the $2Jp \times 2Jp$ matrices $\mathbf{A}_k = \operatorname{diag}(\mathbf{e}_k^\top, \mathbf{e}_k^\top, \ldots, \mathbf{e}_k^\top)$ and

$$\mathbf{B}_k = \begin{pmatrix} \mathbf{D}_k & \mathbf{0}_{p\times 2p} & \cdots & \mathbf{0}_{p\times 2p} \\ \mathbf{0}_{p\times 2p} & \mathbf{D}_k & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0}_{p\times 2p} \\ \mathbf{0}_{p\times 2p} & \cdots & \mathbf{0}_{p\times 2p} & \mathbf{D}_k \end{pmatrix}$$

with a $p \times 2p$ dimensional matrix $\mathbf{D}_k = [\operatorname{diag}(\mathbf{e}_k - J^{-1}\mathbf{1}); \operatorname{diag}(\mathbf{e}_k)]$, where $\mathbf{1} = (1, \ldots, 1)^\top$ is a p-dimensional vector, $\mathbf{e}_k$ is the kth standard basis vector for a Euclidean space with dimension p and $\mathbf{0}_{p\times 2p}$ is the $p \times 2p$ zero matrix with all its entries being zero. Because of this reformulation as a second order cone programming (SOCP) problem, the problem (2.3) can be minimized by any of the many available numerical solvers of SOCP problems. In our simulations and real data analysis, we used the package 'cvx' in MATLAB, see CVX Research, Inc. [7] and Grant and Boyd [18] for details. Note that the dimension of $\mathbf{x}$ is $2Jp$. When p is large, Jp is very large so that the optimization can lead to a grave difficulty of data handling for available softwares. This difficulty generally occurs if one would consider penalized methods based on kernel smoothing in high dimensional models. To circumvent the problem, we used an iterative algorithm to minimize (4.1) in a coefficient(covariate) wise manner for our simulations and in our data example. The idea of coordinatewise optimization is widely used in high dimensional models for similar reasons [see 16, for example]. Although we observed in our simulations that our iterative algorithm converges in a few iterations ($3 \sim 10$, and on average about 4.9 iterations), computation is not fast enough. The reason is that one has to solve a SOCP problem numerically at each covariate-wise step. It deserves further study to develop more efficient and fast computational algorithms.

In our numerical work we used the Epanechnikov kernel $K(u) = 3/4 \cdot (1 - u^2)I(|u| \leq 1)$ with bandwidth $h = 0.15$. To select the regularization parameter λ_1 in (2.2), we used a 5-fold cross validation estimate of the prediction error. For this, we partitioned randomly the original sample into 5 groups of subsamples, $\mathcal{X}_1,\ldots,\mathcal{X}_5$. Then, for each j, the sample with the jth partition removed, $\mathcal{X}_{-j}$, is used for estimation whereas the jth partition, $\mathcal{X}_j$, is used for validation. For the method (2.2), we selected the regularization parameter λ_1 that minimizes the cross validation criterion

$$\sum_{j=1}^{5} \sum_{i \in \mathcal{X}_j} \left[Y_i - \mathbf{X}_i^\top \tilde{\mathbf{m}}_{-j}(Z_i) \right]^2, \tag{4.2}$$

where $\tilde{\mathbf{m}}_{-j}(\cdot)$ is the estimate obtained by applying (2.2) to the data $\mathcal{X}_{-j}$. For selecting the regularization parameter λ_2 in (2.3), we used BIC as defined in (2.5). For the low dimensional case, $(n, p) = (200, 10)$, we used BIC with $C_n = 1$. In the high-dimensional cases $(n, p) = (200, 100)$ and $(200, 200)$, we chose λ_2 for the method (2.3) based on BIC with $C_n = \sqrt{\log p}$.

4.2. Model identification and estimation of penalized methods

We simulated the varying coefficient model in both specifications: in the i.i.d. and in the time series settings, introduced in Section 2. We generated data from the following models:

- Model I (i.i.d. setting):

$$Y_i = \sum_{j=1}^{p} X_i^{(j)} m_j^0(Z_i) + \epsilon_i, \tag{4.3}$$

  where $X_i^{(1)} \equiv 1$. The covariates $(X_i^{(2)}, \ldots, X_i^{(p)})^\top$ are generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = (\sigma_{j_1,j_2})$ with $\sigma_{j_1,j_2} = 0.5^{|j_1 - j_2|}$, the index variables Z_i are from a uniform distribution $U[0, 1]$, the random errors ϵ_i are from $N(0, 0.5^2)$ and for the coefficient functions we choose

$$m_1^0(z) = 2\sin(2\pi z),\ m_2^0(z) = 4z(1 - z),\ m_4^0(z) = 0.5,\ m_5^0(z) = 0.5,$$

  and $m_j^0(z) = 0$ for $j = 3$ and $j \geq 6$.
- Model II (time series setting):

$$Y_i = 0.4\sin(2\pi i/n)Y_{i-1} + 4i/n(1 - i/n)X_i^{(1)} + 0.5X_i^{(2)} + \epsilon_i,$$

  where $\{\epsilon_i\}$ and $\{\mathbf{X}_i\}$ are independently generated by the AR(1) models:

$$X_i^{(j)} = 0.5X_{i-1}^{(j)} + W_i^{(j)},\ j = 1, \ldots, p/2 \ \text{ and } \ \epsilon_i = 0.5\epsilon_{i-1} + \eta_i.$$

  Here $W_i^{(j)}$ are independently generated from $N(0, 1)$ and η_i are i.i.d. from $N(0, 0.25^2)$.

TABLE 1
Simulation results in Model I. In Model I the respective numbers of varying, non-varying and zero coefficients are $2, 2$ *and* $p - 4$.

Method	p	ISE	rISE	CM	$N_{V\to V}$	$N_{I\to V}$	$N_{Z\to V}$	$N_{I\to I}$	$N_{V\to I}$	$N_{Z\to I}$
LASSO	10	0.0775	2.3190	0.00	2.00	2.00	3.80	0.00	0.00	0.00
SCAD		0.0383	1.0304	0.90	2.00	0.07	0.00	1.93	0.00	0.03
LASSO	100	0.1111	3.4336	0.00	2.00	2.00	14.79	0.00	0.00	0.00
SCAD		0.0397	1.0931	0.81	1.98	0.00	0.00	2.00	0.02	0.17
LASSO	200	0.1195	3.3620	0.00	2.00	2.00	19.53	0.00	0.00	0.00
SCAD		0.0419	1.1417	0.77	1.98	0.00	0.00	2.00	0.02	0.21

TABLE 2
Simulation results in Model II. In Model II the respective numbers of varying, non-varying and zero coefficients are $2, 1$ *and* $p - 3$.

Method	p	ISE	rISE	CM	$N_{V\to V}$	$N_{I\to V}$	$N_{Z\to V}$	$N_{I\to I}$	$N_{V\to I}$	$N_{Z\to I}$
LASSO	10	0.0291	2.3126	0.00	2.00	1.00	2.02	0.00	0.00	0.00
SCAD		0.0140	1.0079	0.83	2.00	0.15	0.00	0.85	0.00	0.02
LASSO	100	0.0342	2.5940	0.00	2.00	1.00	12.56	0.00	0.00	0.00
SCAD		0.0145	1.0281	0.90	2.00	0.01	0.00	0.99	0.00	0.10
LASSO	200	0.0346	2.6620	0.00	2.00	1.00	14.78	0.00	0.00	0.00
SCAD		0.0140	1.0397	0.83	2.00	0.00	0.00	1.00	0.00	0.17

In both models, we took $n = 200$ and $p = 10, 100, 200$ in order to see the empirical performance of the methods when the number of the variables varies with the sample size. In the time series scenario, i.e., Model II, we considered $p/2$ lags of the response variable $Y_{i-1}, \ldots, Y_{i-p/2}$ along with $X_i^{(j)}$, $j = 1, ..., p/2$ as potential predictors.

For an assessment of the model selection, we computed the proportion (CM) how the true semiparametric model was correctly selected out of 100 Monte Carlo replications, that is, the proportion of cases where the procedure correctly identified both the true index sets (V^0, I^0). We also report the number of correct and incorrect identifications of the varying and non-varying coefficient functions: $(N_{V\to V})$ denotes the average number of correctly identified varying components, $(N_{I\to V})$ the number of non-varying components classified as varying and $(N_{Z\to V})$ the number of zeros incorrectly identified as varying. Furthermore, $(N_{I\to I})$ is the number of correctly identified non-varying components, $(N_{V\to I})$ the number of varying components classified as non-varying, and $(N_{Z\to I})$ the number of zeros incorrectly identified as non-varying. As measures of estimation accuracy we report the average of the integrated squared error (ISE), $\sum_{j=1}^{p} \int (m_j(z) - m_j^0(z))^2 dz$, and the median value of the relative integrated squared error with respect to the oracle estimator (rISE). As above the oracle estimator is defined as the minimizer of (3.7) subject to the knowledge of the true index sets (V^0, I^0).

Tables 1 and 2 summarize the simulation results of the LASSO-estimator, see (2.2) and the SCAD-estimator, see (2.3), with penalty constant chosen by

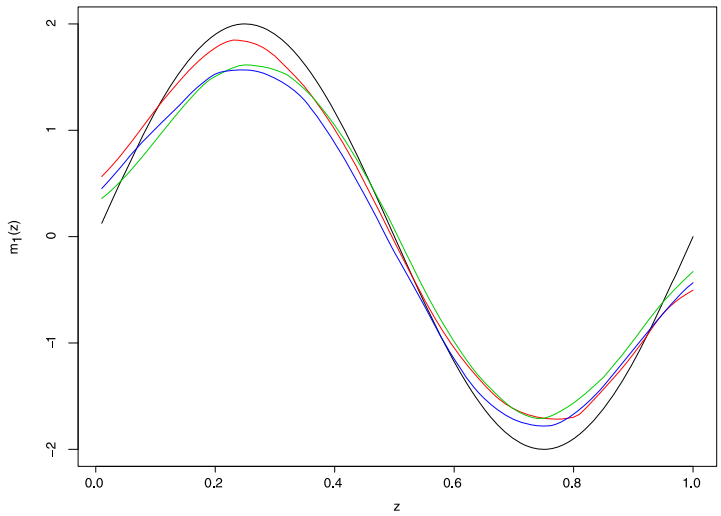

FIG 1. *The first coefficients estimates $\tilde{m}_1$ of the LASSO that shows good/median/poor performance: they correspond to the 0.25 (red), 0.5 (green), 0.75 (blue) quantiles of the ISE results.*

BIC. The values of ISE and rISE show that both methods seem to work in the simulation scenarios. This is expected from Theorems 3.1, 3.4 and 3.5. From the tables, we also see, that the LASSO-estimator is not capable to discriminate non-varying components from varying coefficients: the resulting values of CM, $\mathrm{N}_{I\to I}$, $\mathrm{N}_{V\to I}$ and $\mathrm{N}_{Z\to I}$ are always zero. Furthermore, the table shows that this method tends to include more unnecessary varying components. In contrast, the SCAD-method correctly discriminates both varying and non-varying coefficients from zeros so that it gives a quite accurate estimation, also compared to the oracle estimator. That LASSO performs relatively worse compared to SCAD this might also be caused by the fact that LASSO generally has bias terms for all coefficient functions because penalization applies to all coefficients by the same amount. In Figure 1, the first components, $\tilde{m}_1$, of the LASSO estimates in Model I with $p = 200$ are displayed for the three samples that show good/median/poor performances in estimation. More precisely, the estimates are shown for the random samples corresponding to the 0.25, 0.5, 0.75 quantiles of the ISE results, respectively. From the figure, it can be seen that all the estimates have a bias but that they follow the shape of the true coefficient function m_1^0 quite well. The simulation results confirm the theoretical results in Section 3.

4.3. Consistency of BIC in semiparametric model identification

We carried out additional simulations to see how well the criterion $\mathrm{BIC}(V, I)$ in (2.6) performs in model selection. For s_n in the definition of $\mathcal{M}$, we set $s_n = 20$ in the simulations. As discussed in Section 3.4 it is computationally infeasible to calculate all values of $\mathrm{BIC}(V, I)$ within $\mathcal{M}$ when p is large and

TABLE 3
Model selection and estimation results of BIC(V, I).

	p	ISE	rISE	CM	$N_{V\to V}$	$N_{I\to V}$	$N_{Z\to V}$	$N_{I\to I}$	$N_{V\to I}$	$N_{Z\to I}$
Model I	10	0.0368	1.0000	0.95	2.00	0.03	0.00	1.97	0.00	0.02
	100	0.0396	1.0000	0.90	1.94	0.00	0.00	2.00	0.06	0.04
	200	0.0424	1.0000	0.88	1.91	0.00	0.00	2.00	0.09	0.03
Model II	10	0.0143	1.0000	0.87	2.00	0.11	0.00	0.89	0.00	0.02
	100	0.0136	1.0000	0.93	2.00	0.00	0.00	1.00	0.00	0.07
	200	0.0148	1.0000	0.84	2.00	0.00	0.01	1.00	0.00	0.16

that for this reason, we propose to replace $\mathcal{M}$ by a subset of $\mathcal{M}$ given by the regularization path of the penalization method (2.3). That means we let BIC only run over the sets $(\hat{V}_{\lambda_2}, \hat{I}_{\lambda_2})$ with $|\hat{A}_{\lambda_2}| \leq s_n$ and $\lambda_2 > 0$. We called the unpenalized estimator corresponding to the choice of $(\hat{V}_{\lambda_2}, \hat{I}_{\lambda_2})$ the BIC estimator in Section 3.4. Clearly, the unpenalized estimator with the choice $(\hat{V}_{\lambda_2}, \hat{I}_{\lambda_2})$ is not equal to the penalized estimator $(\hat{\mathbf{m}}_{\lambda_2}, \hat{\mathbf{m}}^{(1)}_{\lambda_2})$ with the same value of λ_2. Thus BIC$(\hat{V}_{\lambda_2}, \hat{I}_{\lambda_2})$ is not equal to BIC(λ_2). Table 3 shows the results of model selection by using the BIC(V, I)-criterion over the described subset of $\mathcal{M}$ and it also gives the values of the integrated squared error for the BIC-estimator. The table shows that in our simulations model choice by BIC(V, I) works pretty well and leads to a very accurate estimator. However, the differences between the SCAD-estimator and the BIC-estimator are small. They both show a very excellent performance, in particular compared to the LASSO-estimator in our simulations.

5. A data example

In this section, we apply our methods to daily observations of NASDAQ composite index data from January 1, 1998 to December 31, 2011 ($n = 3523$). The data include daily returns R_i, i.e., the differences between closing logarithmic prices from today and yesterday, and the high-low ranges Y_i, i.e., the differences between the highest and lowest logarithmic prices of a day. The latter has been proposed as a measure of daily volatility in finance. Figure 2 shows the time series plots of the data. Note that the period of the data includes striking financial crisis events: (i) the internet bubble burst in March, 2000 and the aftermath continued until 2002; (ii) the largest bankruptcy in U. S. history (the collapse of Lehman Brothers) occurred in September, 2008.

The data plots show changes over time in the time series dynamics. In particular, one sees pattern in the conditional variance as heteroskedasticity and volatility clustering. This motivates the use of time varying coefficient models. We have fitted such a model with the daily volatility Y_i as response variable. For (potential) covariates, we took the high-low ranges as well as the value, the squared value, the sign, the negative part and the squared negative part of the daily returns. All these values have been taken from the last 4 weeks = 20 work-

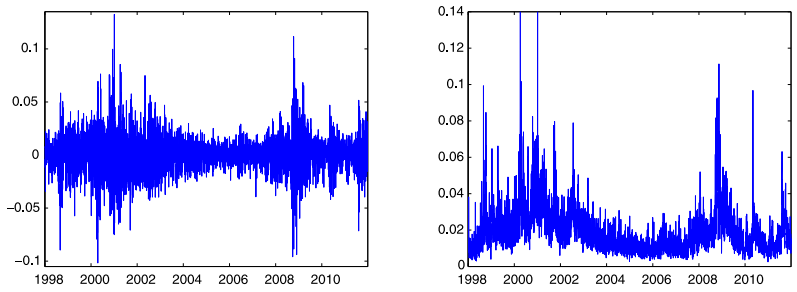

FIG 2. *Time series plots for the daily return R_i (left) and the high-low range Y_i (right).*

ing days. The latter terms are included in the model to check for asymmetric pattern. Thus, we have 120 covariates (except intercept) and the model is given by

$$Y_i = m_0\left(\frac{i}{n}\right) + \sum_{j=1}^{120} X_i^{(j)} m_j\left(\frac{i}{n}\right) + \epsilon_i, i = 1, \ldots, 3523, \tag{5.1}$$

with $\mathbf{X} = (Y_{i-1}, \ldots, Y_{i-20}, R_{i-1}, \ldots, R_{i-20}, R_{i-1}^2, \ldots, R_{i-20}^2, I(R_{i-1} < 0), \ldots, I(R_{i-20} < 0), R_{i-1}I(R_{i-1} < 0), \ldots, R_{i-20}I(R_{i-20} < 0), R_{i-1}^2 I(R_{i-1} < 0), \ldots, R_{i-20}^2 I(R_{i-20} < 0))^\top$. In the analysis the variables are standardized to have zero mean and unit variance, although all results are presented on the original scale of the data.

We applied the penalization methods (2.2) and (2.3) to the dataset. The Epanechnikov kernel was used with a bandwidth that spans approximately one year and a half. As in the simulations we chose the regularization parameter λ_1 for (2.2) by cross validation and the choice of λ_2 for (2.3) is based on ordinary and high dimensional BIC. Method (2.2) identified 55 nonzero coefficient functions whereas method (2.3) with both versions of BIC selected 12 nonzeros, among them, 5 varying and 7 non-varying components, see Table 4 and Table 5. Table 6 contains the estimates of the coefficients in the data example that were classified as non-varying by the (2.3). Figure 3 shows the plots of the estimated (nonconstant) coefficient functions. The model fit makes sense. First, this holds for the signs of the selected coefficients. This also concerns the selected covariates depending on daily returns: $R_{i-1}I(R_{i-1} < 0), R_{i-2}I(R_{i-2} < 0)$ and $R_{i-1}^2 I(R_{i-1}^2 < 0)$. This choice implies an asymmetric effect of returns on volatility, which is well documented in the literature. Furthermore, in Figure 3 one sees that during the financial crisis periods the daily volatility tends to react more strongly to the volatilities and the (negative) returns of last days. However, the curves differ in their shape. A past return (volatility) seems more (less) influential in increasing volatility in the first financial crisis period (i) than in the second period (ii). This may be explained by the difference in the pattern of Y_i during the two financial crisis periods: while rather sporadic peaks and drop-offs were observed during the whole period (i), a number of peaks tend to be concentrated within a relatively narrow time span (late 2008) during period

TABLE 4
The selected covariates in the data example by (2.2).

Coefficient functions corresponding to	Identified as
intercept, Y_{t-k}, $k = 1, ..., 6, 8, ..., 11, 13, ...17, 19, 20$, R_{i-k}, $k = 1, ..., 5, 9, ..., 12, 19$, R^2_{i-k}, $k = 4, 6, 9, 11, 12, 19$, $I(R_{i-k} < 0)$, $k = 4, 5, 9, 12$, $R_{i-k}I(R_{i-k} < 0)$, $k = 1, ...5, 8, ...11$, $R^2_{i-k}I(R_{i-k} < 0)$, $k = 1, 4, 5, 6, 8, 10, 18, 19$	nonzero (varying)

TABLE 5
The selected covariates in the data example by (2.3).

Coefficient functions corresponding to	Identified as
intercept, Y_{t-1}, Y_{t-3}, $R_{i-1}I(R_{i-1} < 0)$, $R^2_{i-1}I(R_{i-1} < 0)$	varying
Y_{t-2}, Y_{t-4}, Y_{t-5}, Y_{t-6},Y_{t-8}, Y_{t-10}, $R_{i-2}I(R_{i-2} < 0)$	non-varying

TABLE 6
Estimates of the coefficients in the data example that were classified as non-varying.

Y_{t-2}	Y_{t-4}	Y_{t-5}	Y_{t-6}	Y_{t-8}	Y_{t-10}	$R_{i-2}I(R_{i-2} < 0)$
0.1070	0.0520	0.0155	0.0248	0.0202	0.0164	-0.0285

(ii). This suggests that past volatilities can predict daily volatilities better in period (ii) than in (i).

Following a referee's suggestion, we checked prediction/one-day-ahead forecasts for the model chosen by LASSO and for the model chosen by SCAD. For this purpose, we used the selected sets $\hat{V}$ and $\hat{I}$ of varying and non-varying coefficients via either (2.2) or (2.3) as listed in Table 4 or 5, respectively. With these two choices of $(\hat{V}, \hat{I})$, we fitted the time varying coefficient model to the observations up to the time r, $\{\mathbf{X}_i, Y_i : i = 1, \ldots, r\}$ with $r_0 \le r \le n$ where $r_0 = 3000$. Our one-day-ahead forecasts $\hat{Y}_{r+1}$ are given as $\mathbf{X}^\top_{r+1}\bar{\mathbf{m}}_{\hat{V},\hat{I}}(1; r)$, where $\bar{\mathbf{m}}_{\hat{V},\hat{I}}(1; r)$ are the estimated coefficients using the r observations $\{\mathbf{X}_i, Y_i : i = 1, \ldots, r\}$ with rescaled time i/r. The estimates $\bar{\mathbf{m}}_{\hat{V},\hat{I}}(1; r)$ were computed as $\bar{\mathbf{m}}_{\hat{V},\hat{I}}$ (see the discussion after (2.6)) but with the full data set replaced by $\{\mathbf{X}_i, Y_i : i = 1, \ldots, r\}$. The bandwidth h_r was chosen so that the kernel window contained the same number of observations as in the original model with scale i/n. The forecast error $(n-r_0)^{-1}\sum_{r=r_0}^{n-1} |\hat{Y}_{r+1} - Y_{r+1}|$ was 0.0076 for the LASSO (2.2) and 0.0059 for the SCAD (2.3).

6. Conclusion

This paper closes a gap in recent interests in sparse high dimensional nonparametric regression. Most papers in this area were only concerned with sieve and orhtogonal series estimation. In this paper we have developed a penalized estimation method based on kernel smoothing. This has been done for a central model

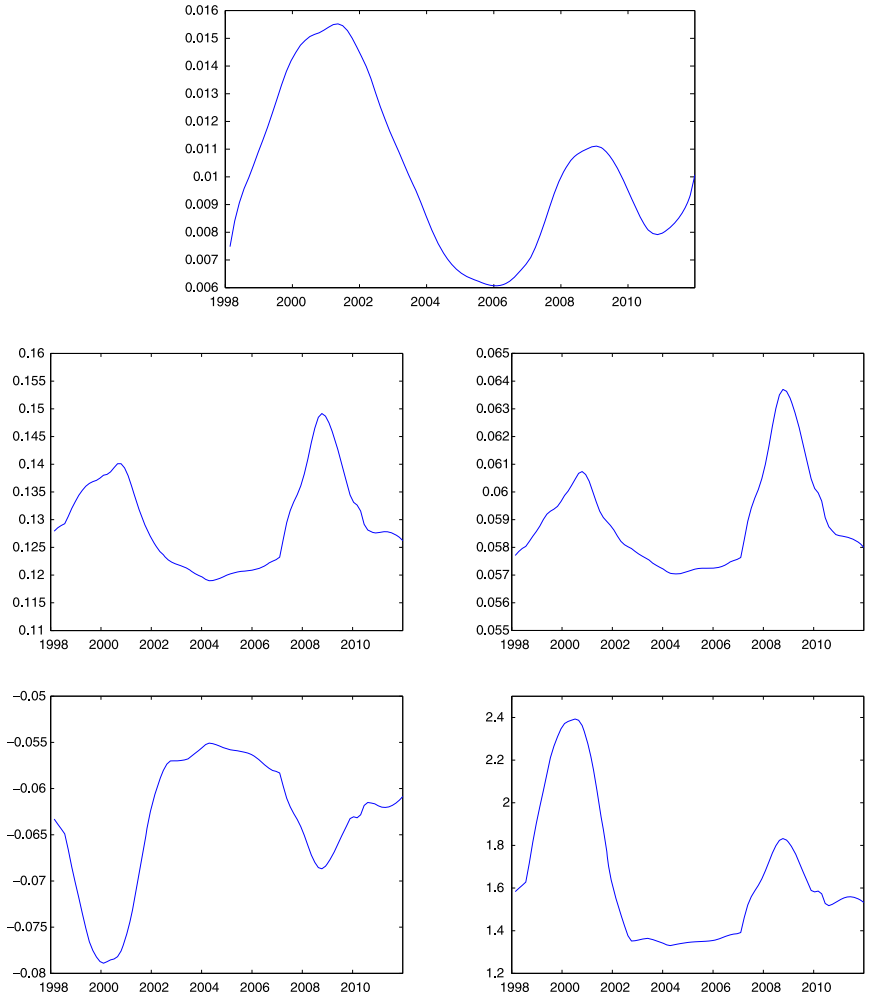

FIG 3. *Estimated coefficient functions corresponding to the intercept (top) and the variables* Y_{i-1} *(middle left),* Y_{i-3} *(middle right),* $R_{i-1}I(R_{i-1} < 0)$ *(bottom left),* $R_{i-1}^2 I(R_{i-1} < 0)$ *(bottom right).*

of sparse high dimensional nonparametric regression. We considered high dimensional varying coefficient models for two settings: for i.i.d. observations and for time varying coefficient models. We showed that our methods can be easily numerically implemented. We proposed several adaptations of group LASSO and SCAD to the local linear kernel method and we carefully investigated their theoretical properties in model structure identification and estimation. We showed that the group LASSO has an estimation error with nearly the same accuracy as if the zero coefficient functions would be known but that typically, it is inconsistent in model selection. Furthermore, the group SCAD estimators have the same

asymptotic properties as when one would know the true structure of a partially linear varying coefficient model. We also argue that the penalized estimators of purely parametric components achieve parametric rates of convergence. This is a stronger advantage than a oracle property as typically shown in the high dimensionality literature. Further we proposed an extension of BIC to select the shrinkage parameter for structure identification. We theoretically justified the proposed BIC-methods by showing their consistency in (semiparametric) model choice.

Appendix

Choose $Z_i, i = 1, \ldots, n$ as i.i.d. copies of $Z \sim f$ where f is the density of Z. Let $C = \sup_{z\in[0,1]} f(z)$ and let $N(z)$ be the number of Z_i's which fall into $[z-h, z+h]$. Note that $N(z)$ follows a binomial distribution with parameters n and $\int_{z-h}^{z+h} f(u)du$. If L increases at a polynomial rate of n, that is, $L = O(n^c)$ for some $c > 0$, then

$$\begin{aligned} P(\sup_{\ell=1,\ldots,L} N(z_\ell) > 3Cnh) &\leq L \max_{\ell=1,\ldots,L} \exp\left(-\frac{2(n\int_{z_\ell-h}^{z_\ell+h} f(z)dz - 3Cnh)^2}{n}\right) \\ &\leq L\exp(-2C^2nh^2) \to 0 \end{aligned} \tag{A.1}$$

as $n \to \infty$. Note that using the Hoeffding's inequality we get that $P(W > M) \leq \exp(-2(np-M)^2/n)$, where $W \sim Bin(n,p)$.

Here, we only present the proofs of Theorems 3.1–3.5 and of Proposition 3.1 in the time series settings. The proofs for the i.i.d. setting can be shown following the lines of the proofs in the time series settings together with the fact (A.1), so that we omit these proofs.

A.1. Proof of Lemma 3.1

For a given p-dimensional vector $\mathbf{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, we let $|\mathbf{a}|_\infty = \sup_{1\leq j\leq p} |a_j|$ be the supremum norm of $\mathbf{a}$. The methods leading to Theorem 2.3. of Dümbgen et al. [11] can be used to derive the following lemma for martingales.

Lemma A.1. *For random variables $\boldsymbol{\xi}_t \in \mathbb{R}^p$, assume that $\boldsymbol{\xi}_t$ is $\mathcal{F}_t$-measurable for an increasing σ-field $\mathcal{F}_t$ with $E(\boldsymbol{\xi}_t|\mathcal{F}_{t-1}) = 0$ and $E(\boldsymbol{\xi}_t^2|\mathcal{F}_{t-1}) < \infty$. Then, there exists a constant $C > 0$ such that*

$$E(|\sum_{t=1}^n \boldsymbol{\xi}_t|_\infty^2) \leq C\log p \sum_{t=1}^n E(|\boldsymbol{\xi}_t|_\infty^2).$$

Now, we prove Lemma 3.1. Define $\boldsymbol{\xi}_i^{(s)}(z) = n^{-1}\epsilon_i \mathbf{X}_i K_h(z - Z_i)((Z_i - z)/h)^s$ for $s = 0, 1$, $i = 1, \ldots, n$ and $z \in [0,1]$. Let $\xi_{i,j}^{(s)}(z)$, $1 \leq j \leq p$ be the j th

component of $\boldsymbol{\xi}_i^{(s)}(z)$. Note that $\sum_{i=1}^n \boldsymbol{\xi}_i^{(s)}(z) = \sum_{i:i/n\in[z-h,z+h]} \boldsymbol{\xi}_i^{(s)}(z)$, and that the number of i's satisfying $z-h \le i/n \le z+h$ is uniformly bounded by Cnh, where the constant C does not depend on z. Then, by applying Lemma A.1, one can take a constant $C>0$, not depending on z, so that

$$E|\sum_{i=1}^n \boldsymbol{\xi}_i^{(0)}(z)|_\infty^2 \le Cd_n^2 \frac{\log p}{nh} \quad \text{and} \quad E|\sum_{i=1}^n \boldsymbol{\xi}_i^{(1)}(z)|_\infty^2 \le Cd_n^2 \frac{\log p}{nh} \tag{A.2}$$

for all $z\in[0,1]$. Note that

$$\begin{aligned} &|\int \sum_{i=1}^n [\boldsymbol{\xi}_i^{(0)}(z)^\top \mathbf{m}(z) + \boldsymbol{\xi}_i^{(1)}(z)^\top \mathbf{m}^{(1)}(z)\, h]dz| \\ &\quad \le \sum_{j=1}^p (\|\sum_{i=1}^n \xi_{i,j}^{(0)}\| \|m_j\| + \|\sum_{i=1}^n \xi_{i,j}^{(1)}\| \|m_j^{(1)}\| h). \end{aligned}$$

This implies that $\mathcal{T}_1^c \subset \{\int |\sum_{i=1}^n \boldsymbol{\xi}_i^{(0)}(z)|_\infty^2 dz > \lambda_0^2\} \cup \{\int |\sum_{i=1}^n \boldsymbol{\xi}_i^{(1)}(z)|_\infty^2 dz > \lambda_0^2\}$. Because of (A.2) this implies Lemma 3.1. □

A.2. Proof of Theorem 3.1

Let $\epsilon_i^0(\cdot) = Y_i - \mathbf{X}_i^\top[\mathbf{m}^0(\cdot) + (\mathbf{m}^0)^{(1)}(\cdot)(Z_i - \cdot)]$. Then, observe that on $\mathcal{T} = \mathcal{T}_1 \cap \mathcal{T}_2$,

$$\begin{aligned} &n^{-1}\int \sum_{i=1}^n \epsilon_i^0(z)\mathbf{X}_i^\top(\mathbf{m}(z) + \mathbf{m}^{(1)}(z)(Z_i - z))K_h(z - Z_i)dz \\ &\le (\lambda_0 + C_1C_2a^0h^2)\sum_{j=1}^p (\|m_j\| + h\|m_j^{(1)}\|) \\ &\le \sqrt{2}(\lambda_0 + C_1C_2a^0h^2)P(\mathbf{m}, \mathbf{m}^{(1)}), \end{aligned} \tag{A.3}$$

where C_1 and C_2 are the constants in the assumption (A3) and in the definition of $\mathcal{T}_2$, respectively. From this and the inequality that $L(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^{(1)}) + \lambda_1 P(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^{(1)}) \le L(\mathbf{m}^0, (\mathbf{m}^0)^{(1)}) + \lambda_1 P(\mathbf{m}^0, (\mathbf{m}^0)^{(1)})$, one has that on $\mathcal{T}$,

$$\begin{aligned} &S\big(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}\big) + \lambda_1 P(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^{(1)}) \\ &\quad \le 2\sqrt{2}(\lambda_0 + C_1C_2a^0h^2)P(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}) + \lambda_1 P(\mathbf{m}^0, (\mathbf{m}^0)^{(1)}), \end{aligned}$$

where $\boldsymbol{\Delta}(\cdot) = (\Delta_1(\cdot), \ldots, \Delta_p(\cdot))^\top$ with $\Delta_j(\cdot) = \tilde{m}_j(\cdot) - m_j^0(\cdot)$, $1 \le j \le p$ and $\boldsymbol{\Delta}^{(1)}(\cdot) = (\Delta_1^{(1)}(\cdot), \ldots, \Delta_p^{(1)}(\cdot))^\top$ with $\Delta_j^{(1)}(\cdot) = \tilde{m}_j^{(1)}(\cdot) - (m_j^0)^{(1)}(\cdot)$, $1 \le j \le p$. This gives

$$S\big(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}\big) + \frac{1}{2}\lambda_1 P_{(A^0)^c}(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}) \le \frac{3}{2}\lambda_1 P_{A^0}(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}). \tag{A.4}$$

We conclude that

$$
\begin{aligned}
2S\big(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}\big) + \lambda_1 P(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}) &\leq 4\lambda_1 P_{A^0}(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}) \\
&\leq 4\lambda_1 (a^0)^{1/2} \left(\sum_{j \in A^0} \|\Delta_j\|^2 + h^2 \|\Delta_j^{(1)}\|^2 \right)^{1/2} \\
&\leq 4\lambda_1 (a^0)^{1/2} \phi_n^{-1} \left[S\big(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}\big) \right]^{1/2} \\
&\leq 4\lambda_1^2 a^0 \phi_n^{-2} + S\big(\boldsymbol{\Delta}, \boldsymbol{\Delta}^{(1)}\big).
\end{aligned}
$$

This concludes the proof of Theorem 3.1. Here, the last inequality uses the fact that $2uv \leq u^2 + v^2$ for $u, v \in \mathbb{R}$ and the last second inequality follows directly from (A.4) and the assumption (A4). □

A.3. Proof of Theorem 3.2

Let $r_n = d_n^2 (\log n + \log p)^{1/2} (nh)^{-1/2}$ and $\hat{\Psi}_{jk,s}(z) = n^{-1} \sum_{i=1}^n X_i^{(j)} X_i^{(k)} ((Z_i - z)/h)^s K_h(z - Z_i)$ for $s = 0, 1, 2$. The following lemma is taken from Liebscher [25]. It states an exponential inequality for sums of α-mixing random variables. We will use the result in the proof of Theorem 3.2.

Lemma A.2. (Liebscher, Theorem 2.1) For a triangular array $\xi_{i,n}$, $1 \leq i \leq n$ with α-mixing coefficients $\alpha(k)$, assume that $E\xi_{i,n} = 0$ and $|\xi_{i,n}| \leq b_n < \infty$ a.s. for $1 \leq i \leq n$. Then, for all $1 \leq m \leq n$ and $\epsilon > 4mb_n$,

$$
P(|\sum_{i=1}^n \xi_{i,n}| > \epsilon) \leq 4 \exp\left(-\frac{\epsilon^2}{64 S_m^2 n/m + 8\epsilon m b_n/3} \right) + 4\frac{n}{m}\alpha(m)
$$

where $S_m^2 = \sup_{0 \leq j \leq n-1} E(\sum_{i=j+1}^{\min\{(j+m),n\}} \xi_{i,n})^2$.

Lemma A.3. *Suppose that the assumptions (A1)–(A2), (A5)–(A7) hold. Then, for sufficiently large $M > 0$ and $s = 0, 1, 2$,*

$$
\lim_{n \to \infty} P\left(\sup_{1 \leq j,k \leq p} \sup_{z \in [0,1]} |\hat{\Psi}_{jk,s}(z) - E\hat{\Psi}_{jk,s}(z)| \geq M r_n \right) = 0 \tag{A.5}
$$

so that $\sup_{1 \leq j,k \leq p} \sup_{z \in [0,1]} |\hat{\Psi}_{jk,s}(z) - E\hat{\Psi}_{jk,s}(z)| = O_p(r_n)$ *for* $s = 0, 1, 2$.

Proof. Put $\hat{\Psi}_{jk}(z) = \hat{\Psi}_{jk,0}(z)$. From Lemma 2.2 in Liebscher [25], note that

$$
\begin{aligned}
&E\left(\sum_{i=l+1}^{\min\{(l+m),n\}} \left\{ X_i^{(j)} X_i^{(k)} K\left(\frac{z - Z_i}{h} \right) - E\left[X_i^{(j)} X_i^{(k)} K\left(\frac{z - Z_i}{h} \right) \right] \right\} \right)^2 \\
&\quad \leq C_1 m d_n^4
\end{aligned}
$$

for some constant $0 < C_1 < \infty$. Applying Lemma A.2 with $m = r_n^{-1} d_n^2$, we get that for $1 \leq j, k \leq p$ and $z \in [0, 1]$

$$
\begin{aligned}
&P\Big(|\hat{\Psi}_{jk}(z) - E\hat{\Psi}_{jk}(z)| > Mr_n\Big) \\
&= P\left(\left|\sum_{i=1}^{n} X_i^{(j)}X_i^{(k)}K\Big(\frac{z-Z_i}{h}\Big) - E(X_i^{(j)}X_i^{(k)})K\Big(\frac{z-Z_i}{h}\Big)\right| > Mr_n nh\right) \\
&\le C_2\left[\exp\left(-C_2^{-1}\frac{(Mr_nnh)^2}{nhd_n^4 + Mr_nnh\cdot r_n^{-1}d_n^2\cdot d_n^2}\right) + nh(r_n^{-1}d_n^2)^{-1-\alpha}\right] \\
&\le C_2\left[\exp\left(-(2C_2)^{-1}M\, r_n^2\,\frac{nh}{d_n^4}\right) + nhr_n^{1+\alpha}d_n^{-2(1+\alpha)}\right] \qquad \text{(A.6)}
\end{aligned}
$$

for $M > 1$ and some $0 < C_2 < \infty$.

Let $\delta_n = r_n h$ and $B_\ell \equiv \{z : |z - z_\ell| \le \delta_n\}, 1 \le \ell \le N$ be minimal number of balls with radius δ_n that cover $[0,1]$. By Lipschitz continuity of K, observe that for all $z \in B_\ell$,

$$
|\hat{\Psi}_{jk}(z) - \hat{\Psi}_{jk}(z_\ell)| \le r_n(nh)^{-1}\sum_{i=1}^{n}|X_i^{(j)}X_i^{(k)}|\tilde{K}\Big(\frac{z_\ell - z}{h}\Big)
$$

for some bounded and nonnegative function $\tilde{K}$ with compact support. Then,

$$
\begin{aligned}
\sup_{z\in B_\ell}|\hat{\Psi}_{jk}(z) - E\hat{\Psi}_{jk}(z)| &\le |\hat{\Psi}_{jk}(z_\ell) - E\hat{\Psi}_{jk}(z_\ell)| \\
&\quad + r_n\Big(|\tilde{\Psi}_{jk}(z_\ell)| + E|\tilde{\Psi}_{jk}(z_\ell)|\Big) \\
&\le |\hat{\Psi}_{jk}(z_\ell) - E\hat{\Psi}_{jk}(z_\ell)| \\
&\quad + |\tilde{\Psi}_{jk}(z_\ell) - E\tilde{\Psi}_{jk}(z_\ell)| + 2r_nM
\end{aligned}
$$

for $r_n < 1$ and sufficiently large $M > 0$, where $\tilde{\Psi}_{jk}(z) = n^{-1}\sum_{i=1}^{n}|X_i^{(j)}X_i^{(k)}| \times\tilde{K}_h(z - Z_i)$. From this, (A.6) and the fact that $N \le (1 + 4\delta_n^{-1})$ (see Section 2.4 in van de Geer [31]), we have

$$
\begin{aligned}
&P\Big(\sup_{1\le j,k\le p}\sup_{z\in[0,1]}|\hat{\Psi}_{jk}(z) - E\hat{\Psi}_{jk}(z)| > 4Mr_n\Big) \\
&\le p^2N\max_{\substack{1\le j,k\le p\\ \ell=1,\dots,N}}\Big[P\Big(|\hat{\Psi}_{jk}(z_\ell) - E\hat{\Psi}_{jk}(z_\ell)| > Mr_n\Big) \\
&\qquad + P\Big(|\tilde{\Psi}_{jk}(z_\ell) - E\tilde{\Psi}_{jk}(z_\ell)| > Mr_n\Big)\Big] \\
&\le C_3\left(p^2(r_nh)^{-1}(np)^{-M/(2C_2)} + p^2nr_n^{\alpha}d_n^{-2(1+\alpha)}\right) \to 0 \quad \text{as} \quad n\to\infty
\end{aligned}
$$

for sufficiently large $M > 0$, which completes the proof when $s = 0$. The fact (A.5) with $s = 1, 2$, can be proved along the lines of the proof with $s = 0$. □

Let $S^*\big(\mathbf{m}, \mathbf{m}^{(1)}\big) = \int \mathbf{m}(z)^\top\mathbf{\Sigma}(z)\mathbf{m}(z)dz + h^2\int \mathbf{m}^{(1)}(z)^\top\mathbf{\Sigma}(z)\mathbf{m}^{(1)}(z)dz$. Given $z \in [0,1]$ and $s = 0, 1, 2$, define $\hat{\mathbf{\Psi}}_s(z)$ to be the $p\times p$ matrix whose (j,k)th element equals $\hat{\Psi}_{jk,s}(z)$. Lemma A.3 shows that there exists a constant $M_1 > 0$ such that $\sup_{1\le j,k\le p}\sup_{z\in[0,1]}|\hat{\Psi}_{jk,s}(z) - E\hat{\Psi}_{jk,s}(z)| \le M_1r_n$ for $s = 0, 1, 2$ with

probability tending to 1 as $n \to \infty$. Define $\mathcal{E} = \{\sup_{1\le j,k\le p} \sup_{z\in[0,1]} |\hat{\Psi}_{jk,s}(z) - E\hat{\Psi}_{jk,s}(z)| \le M_1 r_n,\ s = 0,1,2\}$. By assumptions (A2) and (A5), it can be proved that for $s = 0, 1, 2$,

$$\begin{aligned} &\sup_{\substack{1\le j,k\le p \\ z\in[0,1]}} \Bigg| n^{-1} \sum_{i=1}^{n} \Sigma_{jk}(i/n) K_h(z - i/n) \left(\frac{i/n - z}{h}\right)^s \\ &\qquad\qquad - \Sigma_{jk}(z) \int_0^1 K_h(z-u) \left(\frac{u-z}{h}\right)^s du \Bigg| \\ &\le M_2 \left(h + \frac{1}{nh^2}\right) \end{aligned} \tag{A.7}$$

for some constant $M_2 > 0$. Let $\mu_s(K; z) = \int_{-1}^{z} u^s K(u) du$ for $s = 0, 1, 2$ and $-1 \le z \le 1$. Since $(\mu_1(K;0))^2 < \mu_0(K;0)\mu_2(K;0)$ by application of the Cauchy-Schwarz inequality, we can take $c_1 > 0$ and $c_2 > 0$ that satisfy $(\mu_1(K;0))^2 \le c_1 c_2$, $c_1 < \mu_0(K;0)$ and $c_2 < \mu_2(K;0)$. Let $\delta = \min\{\mu_0(K;0) - c_1,\ \mu_2(K;0) - c_2\}$ and $M = \max\{M_1, M_2\}$. Note that $|\mu_1(K;\cdot)|$ is decreasing and $\mu_0(K;\cdot)$ and $\mu_2(K;\cdot)$ are increasing in an interval $[0,1]$. By (A.7), we have that on $\mathcal{E}$,

$$\begin{aligned} & S\big(\mathbf{m}, \mathbf{m}^{(1)}\big) \\ = & \int [\mathbf{m}^\top \hat{\mathbf{\Psi}}_0 \mathbf{m} + 2h \mathbf{m}^\top \hat{\mathbf{\Psi}}_1 \mathbf{m}^{(1)} + h^2 (\mathbf{m}^{(1)})^\top \hat{\mathbf{\Psi}}_2 \mathbf{m}^{(1)}] \\ \ge & \ \mu_0(K;0) \int \mathbf{m}(z)^\top \mathbf{\Sigma}(z) \mathbf{m}(z) dz - 2h \left| \mu_1(K;0) \int \mathbf{m}(z)^\top \mathbf{\Sigma}(z) \mathbf{m}^{(1)}(z) dz \right| \\ & + \mu_2(K;0) h^2 \int \mathbf{m}^{(1)}(z)^\top \mathbf{\Sigma}(z) \mathbf{m}^{(1)}(z) dz \\ & - M \left(r_n + h + \frac{1}{nh^2}\right) \int \Big\{ \sum_{j=1}^{p} |m_j(z)| + h |m_j^{(1)}(z)| \Big\}^2 dz \\ \ge & \ \delta S^*\big(\mathbf{m}, \mathbf{m}^{(1)}\big) - 32 M a^0 \left(r_n + h + \frac{1}{nh^2}\right) \sum_{j \in A^0} \big(\|m_j\|^2 + h^2 \|m_j^{(1)}\|^2 \big) \end{aligned}$$

for $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$ satisfying $P_{(A^0)^c}(\mathbf{m}, \mathbf{m}^{(1)}) \le 3 P_{A^0}(\mathbf{m}, \mathbf{m}^{(1)})$. This implies $\phi_n^2 \ge \delta (\phi_n')^2 - 32 M a^0 (r_n + h + (nh^2)^{-1})$ so that $\phi_n^2 \ge (\phi_n')^2 \delta / 2$ with probability tending to one. By similar calculations, it can be proved that there exists $\delta' > 0$ such that $(\phi_n')^2 \ge \delta' (\phi_n)^2$ on a set whose probability tends to one. □

A.4. Proofs of Proposition 3.1 and Theorem 3.3

Let $\varepsilon > 0$ be given. Recall $\hat{\Psi}_{jk,s}(z) = n^{-1} \sum_{i=1}^{n} X_i^{(j)} X_i^{(k)} ((Z_i - z)/h)^s K_h(z - Z_i)$ for $s = 0, 1, 2$, $1 \le j, k \le p$ and $z \in [0,1]$. By Lemma A.3 and (A.7), without

loss of generality, we may assume that for $s = 0, 1, 2$,

$$\hat{\Psi}_{jk,s}(z) - \Sigma_{jk}(z)f(z)\int_0^1 K_h(z-u)\left(\frac{u-z}{h}\right)^s du = o_p(1),$$

uniformly for $z \in [0,1]$ and $1 \le j,k \le p$ so that

$$\hat{\mathbf{S}}_{j,A^0}(z)\hat{\mathbf{S}}^{-1}_{A^0,A^0}(z) = \begin{pmatrix} \mathbf{\Sigma}_{j,A^0}(z)\mathbf{\Sigma}^{-1}_{A^0,A^0}(z) & \mathbf{0}_{1\times a^0} \\ \mathbf{0}_{1\times a^0} & \mathbf{\Sigma}_{j,A^0}(z)\mathbf{\Sigma}^{-1}_{A^0,A^0}(z) \end{pmatrix} + o_p(1), \tag{A.8}$$

where the term $o_p(1)$ is uniform in z and j and $\mathbf{0}_{m\times n}$ denotes the zero matrix with dimension $m\times n$. Define $\tilde{\mathbf{b}}, \tilde{\mathbf{c}}$, denoting $(\tilde{b}_1(\cdot), \ldots, \tilde{b}_p(\cdot))$ and $(\tilde{c}_1(\cdot), \ldots, \tilde{c}_p(\cdot))$, respectively, to be the minimizer of $L(\mathbf{b}, h^{-1}\mathbf{c}) + \lambda_1 P(\mathbf{b}, h^{-1}\mathbf{c})$ with respect to $\mathbf{b}, \mathbf{c}$. Then, $\tilde{b}_j = \tilde{m}_j$ and $\tilde{c}_j = h\tilde{m}_j^{(1)}$, $1 \le j \le p$. Similarly, we put $\mathbf{b}^0 = \mathbf{m}^0$ and $\mathbf{c}^0 = h(\mathbf{m}^0)^{(1)}$.

Suppose $\tilde{A} = A^0$. Then, by KKT condition, observe that

$$\begin{aligned} \left.\frac{\partial L(\mathbf{b}, h^{-1}\mathbf{c})}{\partial b_j}\right|_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}})}(\cdot) + \lambda_1 \tilde{s}_j(\cdot) = 0 \quad \text{a.e.}, \qquad & j \in A^0 \\ \left.\frac{\partial L(\mathbf{b}, h^{-1}\mathbf{c})}{\partial c_j}\right|_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}})}(\cdot) + \lambda_1 \tilde{s}_j^{(1)}(\cdot) = 0 \quad \text{a.e.}, \qquad & j \in A^0 \\ \int \left(\left.\frac{\partial L(\mathbf{b}, h^{-1}\mathbf{c})}{\partial b_j}\right|_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}})}\right)^2 + \left(\left.\frac{\partial L(\mathbf{b}, h^{-1}\mathbf{c})}{\partial c_j}\right|_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}})}\right)^2 \le \lambda_1^2, \qquad & j \notin A^0 \end{aligned} \tag{A.9}$$

where

$$\left.\frac{\partial L(\mathbf{b}, h^{-1}\mathbf{c})}{\partial b_j}\right|_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}})} = -2n^{-1}\sum_{i=1}^n \tilde{e}_i(z)K_h(z-Z_i)X_i^{(j)}$$

and

$$\left.\frac{\partial L(\mathbf{b}, h^{-1}\mathbf{c})}{\partial c_j}\right|_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}})} = -2n^{-1}\sum_{i=1}^n \tilde{e}_i(z)K_h(z-Z_i)X_i^{(j)}(Z_i-z)/h$$

with $\tilde{e}_i(z) = Y_i - \mathbf{X}_i^\top[\tilde{\mathbf{b}}(z) + \tilde{\mathbf{c}}(z)(Z_i-z)/h]$. The first two equations give

$$\begin{aligned} &\mathbf{d}(z) \\ &= \hat{\mathbf{S}}_{A^0,A^0}(z)^{-1}[n^{-1}\sum_{i=1}^n \mathbf{\Gamma}_{i,A^0}(z)e_i^0(z)K_h(z-Z_i) - 2^{-1}\lambda_1(\tilde{\mathbf{s}}_{A^0}(z)^\top\ \tilde{\mathbf{s}}_{A^0}^{(1)}(z)^\top)^\top] \end{aligned}$$

(a.e), where $\mathbf{d}(\cdot) = (\tilde{b}_j(\cdot) - b_j^0(\cdot),\ j \in A^0; \tilde{c}_j(\cdot) - c_j^0(\cdot),\ j \in A^0)^\top$ is a $2a^0$ dimensional vector. Substituting this in the inequality (A.9), one gets

$$\sqrt{\int [2\delta_{1,j} - \lambda_1 \mathbf{\Sigma}_{j,A^0}\mathbf{\Sigma}^{-1}_{A^0,A^0}\tilde{\mathbf{s}}_{A^0}]^2 + [2\delta_{2,j} - \lambda_1 \mathbf{\Sigma}_{j,A^0}\mathbf{\Sigma}^{-1}_{A^0,A^0}\tilde{\mathbf{s}}^{(1)}_{A^0}]^2} \le (1+\epsilon)\lambda_1$$

from (A.8) and the fact that $\tilde{b}_j = \tilde{c}_j = 0$ for $j \notin A^0$. This concludes the proof of Proposition 3.1 because of

$$\sqrt{\int h_1^2 + h_2^2} \leq \sqrt{\int (g_1 - h_1)^2 + (g_2 - h_2)^2} + \sqrt{\int g_1^2 + g_2^2}.$$

We now come to the proof of Theorem 3.3. For $0 \leq z \leq 1$, let $\mathbf{W}(z) = \text{diag}(K_h(Z_1 - z), \ldots, K_h(Z_n - z)) \in \mathbb{R}^{n \times n}$, $\mathbb{D}(z) = (\mathbf{\Gamma}_1(z), \ldots, \mathbf{\Gamma}_n(z))^\top \in \mathbb{R}^{n \times 2p}$, $\mathbb{D}_j(z) = (\mathbf{\Gamma}_{1,j}(z), \ldots, \mathbf{\Gamma}_{n,j}(z))^\top \in \mathbb{R}^{n \times 2}$ and $\mathbb{D}_{A^0}(z) = (\mathbf{\Gamma}_{1,A^0}(z), \ldots, \mathbf{\Gamma}_{n,A^0}(z))^\top \in \mathbb{R}^{n \times 2a^0}$. Denote the approximation error (or bias) $\mathbf{X}^\top[\mathbf{m}^0(Z_i) - \mathbf{m}^0(\cdot) - (\mathbf{m}^0)^{(1)}(\cdot)(Z_i - \cdot)]$ of the conditional mean $E(Y_i|\mathbf{X}_i, Z_i) = \mathbf{X}_i^\top \mathbf{m}^0(Z_i)$ by $B_i(\cdot)$. Also $\mathbf{B} = (B_1, \ldots, B_n)^\top$ and $\mathbf{P} = \mathbf{W}^{1/2}\mathbb{D}_{A^0}(\mathbb{D}_{A^0}^\top \mathbf{W}\mathbb{D}_{A^0})^{-1}\mathbb{D}_{A^0}^\top \mathbf{W}^{1/2}$. We will use the following decomposition of $e_i^0(z)$: $e_i^0(z) = B_i(z) + \epsilon_i$, that implies $\delta^{(j)}(z) = n^{-1}\, \mathbb{D}_j(z)^\top W^{1/2}(z)(\mathbf{I}_{n\times n} - \mathbf{P}(z))W^{1/2}(z)(\mathbf{B}(z) + \epsilon)$, where $\mathbf{I}_{n\times n}$ is the $n \times n$ dimensional identity matrix. Let e_k, $k = 1, 2$ denote a 2-dimensional vector of which the k th element is 1 and the other zero. From the fact that $\mathbf{I}_{n\times n} - \mathbf{P}(z)$ is an idempotent matrix, (A.8) and the assumption (A5) with $C_2 > C$, one can see that on $\mathcal{T}_2$, $\left|n^{-1}e_k^\top \mathbb{D}_j(z)^\top W^{1/2}(z)(\mathbf{I}_{n\times n} - \mathbf{P}(z))W^{1/2}(z)\mathbf{B}\right| \leq C_1 C_2 a^0 h^2$ for $k = 1, 2$, and $z \in [0, 1]$, and that on $\mathcal{S}_1' \cap \mathcal{S}_1$, $\|n^{-1}\mathbf{1}_2^\top \mathbb{D}_j^\top W^{1/2}(\mathbf{I}_{n\times n} - \mathbf{P})W^{1/2}\epsilon\| \leq \lambda_0 + \phi^{-2}a^0 C_2 \lambda_0'$. This completes the proof of Theorem 3.3. □

A.5. *Proof of Theorem 3.4*

Let $\mathbf{m}^*(\cdot) = (m_1^*(\cdot), \ldots, m_p^*(\cdot))^\top$ and $(\mathbf{m}^*)^{(1)}(\cdot) = ((m_1^*)^{(1)}(\cdot), \ldots, (m_p^*)^{(1)}(\cdot))^\top$ be vectors of p functions whose jth entries, for $j \in A^0$, are equal to $\hat{m}_j^{ora}(\cdot)$ and $(\hat{m}_j^{ora})^{(1)}(\cdot)$, respectively, and where the other entries are zero functions. The next lemma gives a rate of convergence for the oracle estimators.

Lemma A.4. *Under the assumptions (A1)–(A3),(A4′) and (A5)–(A7),*

$$\sum_{j \in A^0} \|m_j^* - m_j^0\| + h\|(m_j^*)^{(1)} - (m_j^0)^{(1)}\| = O_p\left(a^0(\phi_n')^{-2}\left(d_n\sqrt{\frac{\log a^0}{nh}} + a^0 h^2\right)\right).$$

Proof. Define

$$P_n = \sum_{j \in A^0} \|m_j^* - m_j^0\| + h\|(m_j^*)^{(1)} - (m_j^0)^{(1)}\|,$$

$$Q_n^{(\ell)} = \sup_{j \in A^0}\left\{\int \left(n^{-1}\sum_{i=1}^n \epsilon_i X_i^{(j)}\left(\frac{Z_i - z}{h}\right)^\ell K_h(z - Z_i)\right)^2 dz\right\}^{1/2}, \tag{A.10}$$

$$R_n = \sup_{z \in [0,1]} \sup_{j,k \in A^0} n^{-1}\sum_{i=1}^n |X_i^{(j)} X_i^{(k)}| K_h(z - Z_i), \tag{A.11}$$

for $\ell = 0, 1$. From optimality of $\hat{m}_j^{ora}$, $(\hat{m}_j^{ora})^{(1)}$ and Theorem 3.2, we get

$$0 \geq L(\mathbf{m}^*, (\mathbf{m}^*)^{(1)}) - L(\mathbf{m}^0, (\mathbf{m}^0)^{(1)})$$
$$\geq -2(\max\{Q_n^{(0)}, Q_n^{(1)}\} + C_1 R_n a^0 h^2) P_n + (C\phi_n')^2 (2a^0)^{-1} (P_n)^2, \quad \text{(A.12)}$$

where C_1 and C are the constants in assumption (A3) and in Theorem 3.2, respectively. By Lemma A.3, (A.7) and similar calculations as in the proof of Lemma 3.1, it can be shown that

$$R_n = O_p(1), \quad Q_n^{(\ell)} = O_p\left(\sqrt{\frac{\log a^0}{nh}} d_n\right).$$

Together with (A.12) this implies the statement of Lemma A.4. □

Proof of Theorem 3.4. Since m_j^*, $(m_j^*)^{(1)}$ for $j \in A^0$ minimizes (3.7) subject to the constraints: $\|m_j\|_c = 0$ and $(m_j)^{(1)} \equiv 0$ for $j \in A^0 \setminus V^0$, one has that for any $0 \leq z \leq 1$,

$$\sum_{i=1}^n \epsilon_i^*(z) X_i^j (Z_i - z)^\ell K_h(z - Z_i) = 0, \qquad \ell = 0, 1,\ j \in V^0, \quad \text{(A.13)}$$

$$\sum_{i=1}^n \int \epsilon_i^*(z) X_i^j K_h(z - Z_i) dz = 0, \qquad j \in A^0 \setminus V^0, \quad \text{(A.14)}$$

where $\epsilon_i^*(\cdot) = Y_i - \mathbf{X}_i^\top [\mathbf{m}^*(\cdot) + (\mathbf{m}^*)^{(1)}(\cdot)(Z_i - \cdot)]$. For $\mathbf{m}(\cdot) = (m_1(\cdot), \ldots, m_p(\cdot))^\top$ and $\mathbf{m}^{(1)}(\cdot) = (m_1^{(1)}(\cdot), \ldots, m_p^{(1)}(\cdot))^\top$, we define $B(\mathbf{m}, \mathbf{m}^{(1)})$, $B_1(\mathbf{m}, \mathbf{m}^{(1)})$ and $B_2(\mathbf{m}, \mathbf{m}^{(1)})$ as the respective integrals $n^{-1} \int \sum_{i=1}^n \epsilon_i^*(z) \mathbf{X}_i^\top [(\mathbf{m}(z) + \mathbf{m}^{(1)}(z)(Z_i - z)] K_h(z - Z_i) dz$, $n^{-1} \int \sum_{i=1}^n \epsilon_i^*(z) \sum_{j \in A^0 \setminus V^0} X_i^{(j)} [(m_j(z) + m_j^{(1)}(z)(Z_i - z)] K_h(z - Z_i) dz$ and $n^{-1} \int \sum_{i=1}^n \epsilon_i^*(z) \sum_{j \notin A^0} X_i^{(j)} [(m_j(z) + m_j^{(1)}(z)(Z_i - z)] K_h(z - Z_i) dz$, and let $\mathbf{c}(\mathbf{m})(\cdot) = (m_1(\cdot) - \int m_1(z) dz, \ldots, m_p(\cdot) - \int m_p(z) dz)^\top$. From (A.13) and (A.14), we observe that for all $\mathbf{m} = (m_1, \ldots, m_p)^\top$ and $\mathbf{m}^{(1)} = (m_1^{(1)}, \ldots, m_p^{(1)})^\top$,

$$B(\mathbf{m}, \mathbf{m}^{(1)})$$
$$= B_1(\mathbf{c}(\mathbf{m}), \mathbf{m}^{(1)}) + B_2(\mathbf{m}, \mathbf{m}^{(1)})$$
$$\leq T_n \left(\sum_{j \in A^0 \setminus V^0} [\|m_j\|_c + h\|m_j^{(1)}\|] + \sum_{j \notin A^0} [\|m_j\| + h\|m_j^{(1)}\| \right), \quad \text{(A.15)}$$

where $T_n = \max\{Q_n^{(0)}, Q_n^{(1)}\} + R_n(a^0 C_1 h^2 + \sum_{j \in A^0} \|m_j^* - m_j^0\| + h\|(m_j^*)^{(1)} - (m_j^0)^{(1)}\|)$, with constant C_1 chosen as in (A3) and with $Q_n^{(\ell)}$ $(\ell = 0, 1)$ and R_n defined as in (A.10)–(A.11). Then, $T_n = O_p(b_n)$ by Lemma 3.1 and Lemma A.4.

From Corollary 3.1, one has

$$\sup_{1 \leq j \leq p} \|\hat{m}_j - m_j^0\| = O_p(b_n) \quad \text{and} \quad \sup_{1 \leq j \leq p} h\|\hat{m}_j^{(1)} - (m_j^0)^{(1)}\| = O_p(b_n).$$

This together with the assumptions (A8), $\max\{\lambda_2, \lambda_2^*\}/\delta \to 0$ and $\min\{\lambda_2, \lambda_2^*\}/b_n \to \infty$ implies

$$v_j = \lambda_2 I(j \notin A^0), \quad \text{and} \quad w_j = \lambda_2^* I(j \notin V^0), \quad \text{(A.16)}$$

for all $j = 1, \ldots, p$ with probability tending to one as $n \to \infty$. Thus on a set (A.16) for all $j = 1, \ldots, p$ holds,

$$\begin{aligned} 0 \quad &\geq \quad L(\hat{\mathbf{m}}, \hat{\mathbf{m}}^{(1)}) - L(\mathbf{m}^*, (\mathbf{m}^*)^{(1)}) \\ &+ \min\{\lambda_2, \lambda_2^*\} \left(\sum_{j \notin A^0} (\|\hat{m}_j\|^2 + h^2 \|\hat{m}_j^{(1)}\|^2)^{1/2} + \sum_{j \notin V^0} (\|\hat{m}_j\|_c^2 + h^2 \|\hat{m}_j^{(1)}\|^2)^{1/2} \right) \\ &\geq \quad -2B(\hat{\mathbf{m}} - \mathbf{m}^*, \hat{\mathbf{m}}^{(1)} - (\mathbf{m}^*)^{(1)}) \\ &\quad + \frac{\min\{\lambda_2, \lambda_2^*\}}{\sqrt{2}} \left(\sum_{j \notin A^0} \|\hat{m}_j\| + h\|\hat{m}_j^{(1)}\| + \sum_{j \notin V^0} \|\hat{m}_j\|_c + h\|\hat{m}_j^{(1)}\| \right) \\ &\geq \quad (-2T_n + \frac{\min\{\lambda_2, \lambda_2^*\}}{\sqrt{2}}) \left(\sum_{j \notin A^0} \|\hat{m}_j\| + \sum_{j \in A^0 \setminus V^0} \|\hat{m}_j\|_c + \sum_{j \notin V^0} h\|\hat{m}_j^{(1)}\| \right) \end{aligned}$$

because of the facts that $(\hat{\mathbf{m}}, \hat{\mathbf{m}}^{(1)})$ is the minimizer of the criterion (2.3) and that (A.15). This implies (i) and so (ii) of Theorem 3.4 because $T_n = O_p(b_n)$ and $\min\{\lambda_2, \lambda_2^*\}/b_n \to \infty$. □

A.6. Proof of Theorem 3.5

Let $(V, I) \in \mathcal{M}$ be given. For simplicity, we denote the corresponding estimator $(\bar{\mathbf{m}}_{V,I}, \bar{\mathbf{m}}_{V,I}^{(1)})$ as $(\bar{\mathbf{m}}, \bar{\mathbf{m}}^{(1)})$ whenever this may cause no confusion. Following similar calculations as in (A.15), we have that

$$\begin{aligned} & B(\bar{\mathbf{m}} - \mathbf{m}^*, \bar{\mathbf{m}}^{(1)} - (\mathbf{m}^*)^{(1)}) \\ \leq \quad & T_n \Big\{ \sum_{j \in I^0 \cap V} [\|\bar{m}_j\|_c + h\|\bar{m}_j^{(1)}\|] + \sum_{j \in (A^0)^c \cap V} [\|\bar{m}_j\| + h\|\bar{m}_j^{(1)}\| \Big\} \\ & + S_n \sum_{j \in (A^0)^c \cap I} |\bar{m}_j|, \end{aligned} \tag{A.17}$$

since $B(\bar{\mathbf{m}} - \mathbf{m}^*, \bar{\mathbf{m}}^{(1)} - (\mathbf{m}^*)^{(1)}) = B_1(\mathbf{c}(\bar{\mathbf{m}}), \bar{\mathbf{m}}^{(1)}) + B_2(\bar{\mathbf{m}}, \bar{\mathbf{m}}^{(1)})$ by (A.13) and (A.14), where $\mathbf{m}^*$, $(\mathbf{m}^*)^{(1)}$, $B(\cdot, \cdot)$, $B_1(\cdot, \cdot)$, $B_2(\cdot, \cdot)$, T_n and $\epsilon_i^*(\cdot)$ are defined as in Appendix A.5, and we denote $S_n = |n^{-1} \sum_{i=1}^n \int \epsilon_i^*(z) K_h(z - Z_i) dz \, \mathbf{X}_i|_\infty$. It can be proved that

$$T_n^2 = O_p\left((nh)^{-1} \log n\right). \tag{A.18}$$

Without loss of generality, we may assume that

$$\inf_{j \in A^0} \|m_j^*\| > \delta/2 \quad \text{and} \quad \inf_{j \in V^0} \|m_j^*\|_c > \delta/2 \tag{A.19}$$

because $\sum_{j \in A^0} \|m_j^* - m_j^0\| + h\|(m_j^*)^{(1)} - (m_j^0)^{(1)}\| = O_p\left((nh)^{-1/2}\right) = o_p(1)$ holds as can be seen by a similar proof as in the proof of Lemma A.4. Here δ

is the constant in assumption (A8). Since $\mathbf{m}^*$ and $(\mathbf{m}^*)^{(1)}$ are equal to $\bar{\mathbf{m}}_{V^0,I^0}$ and $\bar{\mathbf{m}}^{(1)}_{V^0,I^0}$ respectively, we observe that

$$\begin{aligned}
& L(\bar{\mathbf{m}}, \bar{\mathbf{m}}^{(1)}) - L(\bar{\mathbf{m}}_{V^0,I^0}, \bar{\mathbf{m}}^{(1)}_{V^0,I^0}) \\
= \;& -2B(\bar{\mathbf{m}} - \mathbf{m}^*, \bar{\mathbf{m}}^{(1)} - (\mathbf{m}^*)^{(1)}) + S(\bar{\mathbf{m}} - \mathbf{m}^*, \bar{\mathbf{m}}^{(1)} - (\mathbf{m}^*)^{(1)}) \\
\geq \;& -2T_n \Big\{ \sum_{j \in I^0 \cap V} [\|\bar{m}_j\|_c + h\|\bar{m}_j^{(1)}\|] + \sum_{j \in (A^0)^c \cap V} [\|\bar{m}_j\| + h\|\bar{m}_j^{(1)}\| \Big\} \\
& -2S_n \sum_{j \in (A^0)^c \cap I} |\bar{m}_j| + (\phi'')^2 \sum_{j \in A^0 \cup A} \|\bar{m}_j - m_j^*\|^2 + h^2 \|\bar{m}_j^{(1)} - (m_j^*)^{(1)}\|^2 \\
\geq \;& -\frac{T_n^2}{(\phi'')^2} \cdot \big(2|I^0 \cap V| + 2|(A^0)^c \cap V|\big) \qquad \text{(A.20)} \\
& -\frac{S_n^2}{(\phi'')^2} \cdot 2|(A^0)^c \cap I| + (\phi'')^2 \delta^2 \, d_{\mathrm{FN}}/2^2, \qquad \text{(A.21)}
\end{aligned}$$

where $d_{\mathrm{FN}} \equiv d_{\mathrm{FN}}(V, I) = |V^0 \cap A^c| + |V^0 \cap I| + |I^0 \cap A^c|$. Here, $|I^0 \cap V| + |(A^0)^c \cap V| + |(A^0)^c \cap I|$ and d_{FN} are the numbers of false positives or false negatives, respectively, when the model (V, I) is chosen. The last second inequality follows directly from (A.17) and the assumption (A4"). The last inequality uses (A.19) and the fact that given $a > 0$ and $b \in \mathbb{R}$, $ax^2 - 2bx \geq -b^2/a$ for all $x \in \mathbb{R}$.

It suffices to show that as $n \to \infty$,

$$P\Big(\min_{(V,I): d_{\mathrm{FN}}(V,I) \geq 1} \mathrm{BIC}(V, I) > \mathrm{BIC}(V^0, I^0)\Big) \to 1 \quad \text{and} \qquad \text{(A.22)}$$

$$P\Big(\min_{(V,I) \neq (V^0,I^0): d_{\mathrm{FN}}(V,I) = 0} \mathrm{BIC}(V, I) > \mathrm{BIC}(V^0, I^0)\Big) \to 1. \qquad \text{(A.23)}$$

These two properties imply Theorem 3.5.

First, we prove (A.23). Suppose $(V, I) \in \mathcal{M} : d_{\mathrm{FN}}(V, I) = 0$. In this case, $|V^0 \cap V^c| = 0$, $|(I^0)^c \cap I| = |(A^0)^c \cap I|$ and $|I^0 \cap I^c| \leq |(V^0)^c \cap V|$. Then, from (A.18), (A.20), the assumption (A11)–(A12) and the fact that $\log(1+x) \geq -2|x|$ for all $x : |x| < 1/2$, we have that

$$\begin{aligned}
& \mathrm{BIC}(V, I) - \mathrm{BIC}(V^0, I^0) \\
= \;& \log\left(1 + \frac{L(\bar{\mathbf{m}}, \bar{\mathbf{m}}^{(1)}) - L(\bar{\mathbf{m}}_{V^0,I^0}, \bar{\mathbf{m}}^{(1)}_{V^0,I^0})}{L(\bar{\mathbf{m}}_{V^0,I^0}, \bar{\mathbf{m}}^{(1)}_{V^0,I^0})}\right) \\
& + C_n \left\{ (|V| - |V^0|) \frac{\log nh}{nh} + (|I| - |I^0|) \frac{\log n}{n} \right\} \\
\geq \;& -2M \left\{ \frac{2T_n^2}{(\phi'')^2} |(V^0)^c \cap V| + \frac{2S_n^2}{(\phi'')^2} |(A^0)^c \cap I| \right\} \\
& + C_n \left\{ (|(V^0)^c \cap V| - |V^0 \cap V^c|) \frac{\log nh}{nh} + (|(I^0)^c \cap I| - |I^0 \cap I^c|) \frac{\log n}{n} \right\} \\
\geq \;& |(V^0)^c \cap V| \cdot C_n \frac{\log nh}{nh} (1 + o_p(1)) + |(A^0)^c \cap I| \cdot C_n \frac{\log n}{n} (1 + o_p(1)),
\end{aligned}$$

where we take a constant $M > 0$ such that $L(\mathbf{m}^0, (\mathbf{m}^0)^{(1)}) = n^{-1}\sum_{i=1}^n \epsilon_i^2 \int K_h(z - Z_i)dz + o_p(1) > M^{-1}$. This inequality implies that (A.23) holds as $n \to \infty$.

It remains to prove (A.22). Consider the case where $d_{\mathrm{FN}} \geq 1$. From (A.20) and the assumption (A10), one has that with probability tending to one as $n \to \infty$,

$$L(\bar{\mathbf{m}}, \bar{\mathbf{m}}^{(1)}) - L(\bar{\mathbf{m}}_{V^0,I^0}, \bar{\mathbf{m}}^{(1)}_{V^0,I^0}) \geq (\phi'')^2\delta^2/2^3 > 0$$

for all $(V, I) \in \mathcal{M}$ with $d_{\mathrm{FN}}(V, I) \geq 1$. Then, since $\log(1+x) \geq \min\{0.5x, \log 2\}$ for all $x > 0$, we get that for any $(V, I) \in \mathcal{M}$ with $d_{\mathrm{FN}}(V, I) \geq 1$

$$\begin{aligned}
&\mathrm{BIC}(V, I) - \mathrm{BIC}(V^0, I^0) \\
\geq\ & \min\left\{0.5M\left(\frac{2T_n^2}{(\phi'')^2}|(V^0)^c \cap V| + \frac{2S_n^2}{(\phi'')^2}|(A^0)^c \cap I| + \frac{(\phi'')^2\delta^2\, d_{\mathrm{FN}}}{2^2}\right), \log 2\right\} \\
&+ C_n\left\{(|(V^0)^c \cap V| - |V^0 \cap V^c|)\frac{\log n}{nh} + (|(I^0)^c \cap I| - |I^0 \cap I^c|)\frac{\log n}{n}\right\} \\
\geq\ & \min\left\{|(V^0)^c \cap V|\frac{\log n}{nh}C_n(1 + o_p(1)) + |(A^0)^c \cap I|\frac{\log n}{n}C_n(1 + o_p(1))\right. \\
&+ |V^0 \cap V^c|(\phi'')^2\delta^2/2^2(1 + o_p(1)) + |I^0 \cap A^c|(\phi'')^2\delta^2/2^2(1 + o_p(1)), \\
&\left.\log 2 + o_p(1)\right\} \\
\geq\ & \min\{\phi'')^2\delta^2/2^3, \log 2/2\} > 0
\end{aligned}$$

by (A.18), (A.20) and assumptions (A11)–(A12). This completes the proof. □

A.7. On the assumption (A12)

In this subsection, we show that (A12) holds under some technical conditions. Before showing this, we introduce some notation. Let $\mathbf{X}_{i,A^0} = (X_i^{(j)} : j \in A^0)^\top$, $\mathbf{X}_{i,I^0} = (X_i^{(j)} : j \in I^0)^\top$ and $\mathbf{X}_{i,V^0} = (X_i^{(j)} : j \in V^0)^\top$. Given $z \in [0,1]$, define $\mathbf{m}^0_{A^0}(z) = (m_j^0(z) : j \in A^0)^\top$, $\mathbf{m}^0_{V^0}(z) = (m_j^0(z) : j \in V^0)^\top$, and $\mathbf{m}^0_{I^0} = (m_j^0 : j \in I^0)^\top$. In the same way, we define $\mathbf{m}^*_{A^0}(\cdot)$, $\mathbf{m}^*_{V^0}(\cdot)$, $\mathbf{m}^*_{I^0}$, $(\mathbf{m}^*)^{(1)}_{A^0}(\cdot)$, $(\mathbf{m}^*)^{(1)}_{V^0}(\cdot)$, $(\mathbf{m}^0)^{(1)}_{V^0}(\cdot)$ and $(\mathbf{m}^0)^{(2)}_{V^0}(\cdot)$ where $(\mathbf{m}^0)^{(s)} = ((m_j^0)^{(s)} : 1 \leq j \leq p)^\top$ and $(m_j^0)^{(s)}$ is the sth derivative of m_j^0. For $z \in [0,1]$, we let $\mathbf{\Sigma}_{V^0V^0}(z) = (\Sigma_{jk}(z) : j,k \in V^0) \in \mathbb{R}^{|V^0|\times|V^0|}$, $\mathbf{\Sigma}_{I^0V^0}(z) = (\Sigma_{jk}(z) : j \in I^0, k \in V^0) \in \mathbb{R}^{|I^0|\times|V^0|}$ and $\mathbf{\Sigma}_{\cdot,V^0}(z) = (\Sigma_{jk}(z) : 1 \leq j \leq p,\ k \in V^0) \in \mathbb{R}^{p\times|V^0|}$.

Now we will see that under suitable regularity conditions, the following properties hold:

$$\left|n^{-1}\sum_{i=1}^n \int \epsilon_i K_h(z - Z_i)dz\, \mathbf{X}_i\right|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right) \quad \text{and} \tag{A.24}$$

$$\left|\sum_{i=1}^{n}\int \frac{\mathbf{X}_i\mathbf{X}_{i,A^0}^{\top}}{n}[\mathbf{m}_{A^0}^{0}(Z_i)-\mathbf{m}_{A^0}^{*}(z)-(\mathbf{m}^*)_{A^0}^{(1)}(z)(Z_i-z)]K_h(z-Z_i)dz\right|_{\infty}$$
$$=O_p(n^{-1/2}). \tag{A.25}$$

These claims immediately imply (A12).

First, (A.24) can be easily shown by a similar proof as in Lemma 3.1. Note that the oracle estimator $(\mathbf{m}_{A^0}^{*},(\mathbf{m}^*)_{A^0}^{(1)})$ is a standard nonparametric estimator when taking $(V,I)=(V^0,I^0)$, defined in a similar fashion as in Xia et al. [39]. Using similar arguments as in Xia et al. [39], it can be proved that for the nonparametric parts $j\in V^0$,

$$\sup_{0\le z\le 1}|m_j^{*}(z)-m_j^{0}(z)|=O_p(n^{-2/5}\sqrt{\log n}),$$
$$\sup_{0\le z\le 1}|(m^*)_j^{(1)}(z)-(m^0)_j^{(1)}(z)|=O_p(n^{-1/5}\sqrt{\log n}),$$
$$\sup_{h\le z\le 1-h}|\mathbf{\Sigma}_{V^0V^0}(z)(\mathbf{m}_{V^0}^{*}(z)-\mathbf{m}_{V^0}^{0}(z)-\frac{\mu_2}{2}(\mathbf{m}^0)_{V^0}^{(2)}(z)h^2)$$
$$-n^{-1}\sum_{i=1}^{n}\epsilon_i\mathbf{X}_{i,V^0}K_h(z-Z_i)|_{\infty}=O_p(n^{-1/2}),$$

under suitable regularity conditions including $h\approx n^{-1/5}$ and $m_j^0(u)=m_j^0(z)+(m_j^0)^{(1)}(z)(u-z)+(m_j^0)^{(2)}(z)(u-z)^2/2+O(h^3)$ uniformly in $u,\ z:|u-z|\le h$ for $j\in V^0$ (sufficient smoothness condition of the true coefficient functions), where $\mu_2=\int_0^1 u^2K(u)du$. From this with the facts that

$$\sup_{1\le j,k\le p}\sup_{z\in[0,1]}|\hat{\Psi}_{jk,s}(z)-\Sigma_{jk}(z)\int_0^1 K_h(z-u)\left(\frac{u-z}{h}\right)^s du|=o_p(1),\quad s=0,1,2,$$

(refer to Lemma A.3 and (A.7)) and that $|n^{-1}\sum_{i=1}^{n}\epsilon_i\mathbf{X}_{i,V^0}\int K_h(z-Z_i)dz|_{\infty}=O_p(n^{-1/2})$, one can see

$$|\sum_{i=1}^{n}\int \frac{\mathbf{X}_{i,I^0}\mathbf{X}_{i,V^0}^{\top}}{n}[\mathbf{m}_{V^0}^{0}(Z_i)-\mathbf{m}_{V^0}^{*}(z)-(\mathbf{m}^*)_{V^0}^{(1)}(z)(Z_i-z)]K_h(z-Z_i)dz|_{\infty}$$
$$=|\int_h^{1-h}\mathbf{\Sigma}_{I^0,V^0}(z)[\mathbf{m}_{V^0}^{0}(z)+\frac{\mu_2}{2}(\mathbf{m}^0)_{V^0}^{(2)}(z)h^2-\mathbf{m}_{V^0}^{*}(z)]|_{\infty}+o_p(n^{-1/2})$$
$$=O_p(n^{-1/2}).$$

Recall that $\hat{\Psi}_{jk,s}(z)=n^{-1}\sum_{i=1}^{n}X_i^{(j)}X_i^{(k)}((Z_i-z)/h)^sK_h(z-Z_i)$ for $s=0,1,2$ and $z\in[0,1]$. So, for the parametric constant coefficients $j\in I^0$,

$$|\mathbf{m}_j^{*}-\mathbf{m}_j^{0}|=O_p(n^{-1/2}) \tag{A.26}$$

because the term $n^{-1}\sum_{i=1}^{n}\int\mathbf{X}_{i,I^0}\mathbf{X}_{i,I^0}^{\top}K_h(z-Z_i)dz\ \times(\mathbf{m}_{I^0}^{*}-\mathbf{m}_{I^0}^{0})$ is equal to $n^{-1}\sum_{i=1}^{n}\int\epsilon_iK_h(z-Z_i)dz\ \mathbf{X}_{i,I^0}+n^{-1}\sum_{i=1}^{n}\int\mathbf{X}_{i,I^0}\mathbf{X}_{i,V^0}^{\top}[\mathbf{m}_{V^0}^{0}(Z_i)-\mathbf{m}_{V^0}^{*}(z)-$

$(\mathbf{m}^*)^{(1)}_{V^0}(z)(Z_i - z)]K_h(z - Z_i)dz$ by (A.14). This concludes (A.25):

$$
\begin{aligned}
&|\sum_{i=1}^{n} \int \frac{\mathbf{X}_i \mathbf{X}_{i,A^0}^\top}{n} [\mathbf{m}^0_{A^0}(Z_i) - \mathbf{m}^*_{A^0}(z) - (\mathbf{m}^*)^{(1)}_{A^0}(z)(Z_i - z)] K_h(z - Z_i) dz|_\infty \\
&= |\sum_{i=1}^{n} \int \frac{\mathbf{X}_i \mathbf{X}_{i,V^0}^\top}{n} [\mathbf{m}^0_{V^0}(Z_i) - \mathbf{m}^*_{V^0}(z) - (\mathbf{m}^*)^{(1)}_{V^0}(z)(Z_i - z)] K_h(z - Z_i) dz|_\infty \\
&\quad + O_p(n^{-1/2}) \\
&= |\int_h^{1-h} \mathbf{\Sigma}_{\cdot,V^0}(z)[\mathbf{m}^0_{V^0}(z) + \frac{\mu_2}{2}(\mathbf{m}^0)^{(2)}_{V^0}(z)h^2 - \mathbf{m}^*_{V^0}(z)]|_\infty + O_p(n^{-1/2}) \\
&= O_p(n^{-1/2}).
\end{aligned}
$$

References

[1] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data.* Springer, 2011. MR2807761

[2] Z. . Cai. Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136:163–188, 2007. MR2328589

[3] B. Chen and Y. Hong. Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica*, 80:1157–1183, 2012. MR2963885

[4] J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008. MR2443189

[5] M.-Y. Cheng, T. Honda, J. Li, and H. Peng. Nonparametric independence screening and structural identification for ultra-high dimensional longitudinal data. *Annals of Statistics*, 42:1819–1849, 2014. MR3262469

[6] M.-Y. Cheng, T. Honda, and J.-T. Zhang. Forward variable selection for sparse ultra-high dimensional varying coefficient models. *J. Amer. Statisti. Assoc.*, forthcoming, 2015.

[7] CVX Research, Inc. CVX: Matlab software for disciplined convex programming, version 2.0 beta. `http://cvxr.com/cvx`, Sept. 2012.

[8] R. Dahlhaus. On the Kullback-Leibler information divergence of locally stationary processes. *Stoch. Proc. Appl*, 62:139–168, 1996. MR1388767

[9] R. Dahlhaus. Fitting time series models to nonstationary processes. *Annals of Statistics*, 25:1–37, 1997. MR1429916

[10] R. Dahlhaus, M. H. Neumann, and R. V. Sachs. Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli*, 5:873–906, 1999. MR1715443

[11] L. Dümbgen, S. van de Geer, M. Veraar, and J. Wellner. Nemirovski's inequalities revisited. *Amer. Math. Monthly*, 117:138–160, 2010. MR2590193

[12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001. MR1946581

[13] J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statist. and its Interface*, 1:179–195, 2008. MR2425354

[14] J. Fan, J. Lv, and L. Qi. Sparse high-dimensional models in economics. *Annual Review of Economics*, 3:291–317, 2011.

[15] J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.*, 109:1270–1284, 2014. MR3265696

[16] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software*, 33(1), 2010.

[17] P. Fryzlewicz and S. S. Rao. Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli*, 17:320–346, 2011. MR2797994

[18] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. MR2409077

[19] T. Hu and Y. Xia. Adaptive semi-varying coefficient model selection. *Statistica Sinica*, 22:575–599, 2012. MR2954353

[20] Y. Kim, S. Kwon, and H. Choi. Consistent model selection criteria on high dimensions. *J. Machine Learning Research*, 13:1037–1057, 2012. MR2930632

[21] O. Klopp and M. Pensky. Sparse high-dimensional varying coefficient model:non-asymptotic minimax study. *Annals of Statistics.*, 43:1273–1299, 2015. MR3346703

[22] D. Kong, H. D. Bondell, and Y. Wu. Domain selection for the varying coefficient model via local polynomial regression. *Comput. Statist. and Data Analysis*, 83:236–250, 2015. MR3281808

[23] E. R. Lee, H. Noh, and B. U. Park. Model selection via bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.*, 109:216–229, 2014. MR3180558

[24] H. Lian. Variable selection for high-dimensional generalized varying coefficient models. *Statistica Sinica*, 22:1563–1588, 2012. MR3027099

[25] E. Liebscher. Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stoch. Proc. Appl.*, 65:69–80, 1996. MR1422880

[26] B. U. Park, E. Mammen, Y. K. Lee, and E. R. Lee. Varying coefficient regression models: A review and new developments. *International Statistical Review*, 83:36–64, 2015. MR3341079

[27] P. M. Robinson. Nonparametric estimation of time-varying parameters. In P. Hackl, editor, *Statistical Analysis and Forecasting of Economic Structural Change*, pages 253–264. Springer-Verlag, 1989.

[28] P. M. Robinson. Time-varying nonlinear regression. In P. Hackl and A. H. Westland, editors, *Economic Structure Change Analysis and Forecasting*, pages 179–190. Springer, 1991.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. B*, 58:267–288, 1996. MR1379242

[30] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Elect. J. Statist.*, 3:1360–1392, 2009. MR2576316

[31] S. A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[32] M. Vogt. Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40:2601–2633, 2012. MR3097614

[33] D. Wang and K. B. Kulasekera. Parametric component detection and variable selection in varying-coefficient partially linear models. *Journal of Multivariate Analysis*, 112:117–129, 2012. MR2957290

[34] H. Wang and C. Leng. Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, 102:1418–1429, 2007. MR2411663

[35] H. Wang and Y. Xia. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104:747–757, 2009. MR2541592

[36] H. Wang, R. Li, and C.-L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94:553–568, 2007. MR2410008

[37] H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *J. Royal Stat. Soc. B*, 71:671–683, 2009. MR2749913

[38] F. Wei, J. Huang, and H. Li. Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21:1515–1540, 2011. MR2895107

[39] Y. Xia, W. Zhang, and H. Tong. Efficient estimation for semivarying-coefficient models. *Biometrika*, 91:661–681, 2004. MR2090629

[40] L. Xue and A. Qu. Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*, 13: 1973–1998, 2012. MR2956349

[41] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. B*, B68:49–67, 2006. MR2212574

[42] C. H. Zhang and J. Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Annals of Statistics*, 4:1567–1594, 2008. MR2435448

[43] H. Zhang, G. Cheng, and Y. Liu. Linear or nonlinear? Augomatic structure discovery for partially linear mdoels. *Journal of the American Statistical Association*, 106:1099–1112, 2011. MR2894767

[44] T. Zhang and W. B. Wu. Inference of time-varying regression models. *Annals of Statistics*, 40:1376–1402., 2012. MR3015029

[45] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. MR2279469

[46] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509–1533, 2008. MR2435443