The Annals of Statistics
2015, Vol. 43, No. 3, 1273–1299
DOI: 10.1214/15-AOS1309

SPARSE HIGH-DIMENSIONAL VARYING COEFFICIENT MODEL: NONASYMPTOTIC MINIMAX STUDY

BY OLGA KLOPP AND MARIANNA PENSKY[1]

University Paris Ouest and University of Central Florida

The objective of the present paper is to develop a minimax theory for the varying coefficient model in a nonasymptotic setting. We consider a high-dimensional sparse varying coefficient model where only few of the covariates are present and only some of those covariates are time dependent. Our analysis allows the time-dependent covariates to have different degrees of smoothness and to be spatially inhomogeneous. We develop the minimax lower bounds for the quadratic risk and construct an adaptive estimator which attains those lower bounds within a constant (if all time-dependent covariates are spatially homogeneous) or logarithmic factor of the number of observations.

1. Introduction. One of the fundamental tasks in statistics is to characterize the relationship between a set of covariates and a response variable. In the present paper we study the varying coefficient model which is commonly used for describing time-varying covariate effects. It provides a more flexible approach than the classical linear regression model and is often used to analyze the data measured repeatedly over time.

Since its introduction by Cleveland, Grosse and Shyu [8] and Hastie and Tibshirani [13], many methods for estimation and inference in the varying coefficient model have been developed; see, for example, [11, 14, 19, 36] for the kernel-local polynomial smoothing approach, [15–17] for the polynomial spline approach, [7, 13, 14] for the smoothing spline approach and [12] for a detailed discussion of the existing methods and possible applications. In the last five years, the varying coefficient model received a great deal of attention. For example, Wang et al. [34] proposed a new procedure based on a local rank estimator; Kai et al. [18] introduced a semi-parametric quantile regression procedure and studied an effective variable selection procedure. Lee et al. [22] extended the model to the case when the response is related to the covariates via a link function while Zhu et al. [38] studied the multivariate version of the model. Existing methods typically provide asymptotic evaluation of the precision of the estimation procedure under the assumption that the number of observations tends to infinity and is larger than the dimension of the problem.

Received May 2014; revised January 2015.
[1]Supported in part by NSF Grants DMS-11-06564 and DMS-14-07475.
MSC2010 subject classifications. Primary 62H12, 62J05; secondary 62C20.
Key words and phrases. Varying coefficient model, sparse model, minimax optimality.

Recently few authors consider still asymptotic but high-dimensional approach to the problem. Wei et al. [35] applied group Lasso for variable selection, while Lian [23] used extended Bayesian information criterion. Fan et al. [10] applied nonparametric independence screening. Their results were extended by Lian and Ma [24] to include rank selection in addition to variable selection.

One important aspect that has not been well studied in the existing literature is the nonasymptotic approach to the estimation, prediction and variables selection in the varying coefficient model. Here, we refer to the situation where both the number of unknown parameters and the number of observations are large and the former may be of much higher dimension than latter. The only reference that we are aware of in this setting is the recent paper by Klopp and Pensky [20]. Their method is based on some recent developments in the matrix estimation problem. Some interesting questions arise in this nonasymptotic setting. One of them is the fundamental question of the minimax optimal rates of convergence. The minimax risk characterizes the essential statistical difficulty of the problem. It also captures the interplay between different parameters in the model. To the best of our knowledge, our paper presents the first *nonasymptotic minimax study* of the sparse heterogeneous varying coefficient model.

Modern technologies produce very high-dimensional data sets and hence stimulate an enormous interest in variable selection and estimation under a sparse scenario. In such scenarios, penalization-based methods are particularly attractive. Significant progress has been made in understanding the statistical properties of these methods. For example, many authors have studied the variable selection, estimation and prediction properties of the LASSO in high-dimensional setting; see, for example, [2, 4, 5, 32]. A related LASSO-type procedure is the group-LASSO, where the covariates are assumed to be clustered in groups; see, for example, [1, 6, 25, 27, 28, 37], and references therein.

In the present paper, we also consider the case when the solution is sparse, in particular, only a few of the covariates are present, and only some of them are time dependent. This setup is close to the one studied in a recent paper of Liang [23]. One important difference, however, is that in [23], the estimator is not adaptive to the smoothness of the time-dependent covariates. In addition, Liang [23] assumes that all time-dependent covariates have the same degree of smoothness and are spatially homogeneous. On the contrary, we consider a much more flexible and realistic scenario where the time-dependent covariates possibly have different degrees of smoothness and may be spatially inhomogeneous.

In order to construct a minimax optimal estimator, we introduce the block Lasso which can be viewed as a version of the group LASSO. However, unlike in group LASSO, where the groups occur naturally, the blocks in block LASSO are driven by the need to reduce the variance as it is done, for example, in block thresholding. Note that our estimator does not require the knowledge of which of the covariates are indeed present and which are time dependent. It adapts to sparsity, to heterogeneity of the time-dependent covariates and to their possibly spatial inhomogeneous nature. In order to ensure the optimality, we derive minimax lower bounds

for the risk and show that our estimator attains those bounds within a constant (if all time-dependent covariates are spatially homogeneous) or logarithmic factor of the number of observations. The analysis is carried out under the flexible assumption that the noise variables are sub-Gaussian. In addition, it does not require that the elements of the dictionary are uniformly bounded.

The rest of the paper is organized as follows. Section 1.1 provides formulation of the problem while Section 1.2 lays down a tensor approach to estimation. Section 2 introduces notation and assumptions on the model and provides a discussion of the assumptions. Section 3 describes the block thresholding LASSO estimator, evaluates the nonasymptotic lower and upper bounds for the risk, both in probability and in the mean squared risk sense and ensures optimality of the constructed estimator. Section 4 presents examples of estimation when assumptions of the apper are satisfied. Section 5 contains proofs of the statements formulated in the paper.

1.1. *Formulation of the problem.* Let $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \dots, n$ be sampled independently from the varying coefficient model

$$Y = \mathbf{W}^T \mathbf{f}(t) + \xi. \tag{1.1}$$

Here, $\mathbf{W} \in \mathbb{R}^p$ are random vectors of predictors, $\mathbf{f}(\cdot) = (f_1(\cdot), \dots, f_p(\cdot))^T$ is an unknown vector-valued function of regression coefficients and $t \in [0, 1]$ is a random variable with the unknown density function g. We assume that $\mathbf{W}$ and t are independent. The noise variable ξ is independent of W and t, and is such that $\mathbb{E}(\xi) = 0$.

The goal is to estimate vector function $f(\cdot)$ on the basis of observations $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \dots, n$.

In order to estimate $\mathbf{f}$, following Klopp and Pensky [20], we expand it over a basis $(\phi_l(\cdot))$, $l = 0, 1, \dots, \infty$, in $L_2([0, 1])$ with $\phi_0(t) = 1$. Expansion of the functions $f_j(\cdot)$ over the basis, for any $t \in [0, 1]$, yields

$$f_j(t) = \sum_{l=0}^{L} a_{jl}\phi_l(t) + \rho_j(t) \qquad \text{with } \rho_j(t) = \sum_{l=L+1}^{\infty} a_{jl}\phi_l(t). \tag{1.2}$$

If $\boldsymbol{\phi}(\cdot) = (\phi_0(\cdot), \dots, \phi_L(\cdot))^T$ and $\mathbf{A}_0$ denotes a matrix of coefficients with elements $\mathbf{A}_0^{(l,j)} = a_{jl}$, then relation (1.2) can be rewritten as $\mathbf{f}(t) = \mathbf{A}_0^T \boldsymbol{\phi}(t) + \boldsymbol{\rho}(t)$, where $\boldsymbol{\rho}(t) = (\rho_1(t), \dots, \rho_p(t))^T$. Combining formulas (1.1) and (1.2), we obtain the following model for observations $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \dots, n$:

$$Y_i = \operatorname{Tr}\big(\mathbf{A}_0^T \boldsymbol{\phi}(t_i)\mathbf{W}_i^T\big) + \mathbf{W}_i^T \boldsymbol{\rho}(t_i) + \xi_i, \qquad i = 1, \dots, n. \tag{1.3}$$

Below, we reduce the problem of estimating vector function $\mathbf{f}$ to estimating matrix $\mathbf{A}_0$ of coefficients of $\mathbf{f}$.

1.2. *Tensor approach to estimation.* Denote $\mathbf{a} = \mathrm{Vec}(\mathbf{A}_0)$ and $\mathbf{B}_i = \mathrm{Vec}(\boldsymbol{\phi}(t_i)\mathbf{W}_i^T)$. Note that $\mathbf{B}_i$ is the $p(L+1)$-dimensional vector with components $\phi_l(t_i)\mathbf{W}_i^{(j)}$, $l = 0, \dots, L$, $j = 1, \dots, p$, where $\mathbf{W}_i^{(j)}$ is the jth component of vector $\mathbf{W}_i$. Consider matrix $\mathbf{B} \in \mathbb{R}^{n \times p(L+1)}$ with rows $\mathbf{B}_i^T$, $i = 1, \dots, n$, vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$ and vector $\mathbf{b}$ with components $\mathbf{b}_i = \mathbf{W}_i^T \boldsymbol{\rho}(t_i)$, $i = 1, \dots, n$. Taking into account that

$$\mathrm{Tr}\big(\mathbf{A}^T \boldsymbol{\phi}(t_i)\mathbf{W}_i^T\big) = \mathbf{B}_i^T \,\mathrm{Vec}(\mathbf{A}),$$

we rewrite the varying coefficient model (1.3) in a matrix form

$$\mathbf{Y} = \mathbf{B}\mathbf{a} + \mathbf{b} + \boldsymbol{\xi}. \tag{1.4}$$

In what follows, we denote

$$\boldsymbol{\Omega}_i = \mathbf{W}_i \mathbf{W}_i^T, \qquad \boldsymbol{\Phi}_i = \boldsymbol{\phi}(t_i)\big(\boldsymbol{\phi}(t_i)\big)^T, \qquad \boldsymbol{\Sigma}_i = \boldsymbol{\Omega}_i \otimes \boldsymbol{\Phi}_i, \tag{1.5}$$

where $\boldsymbol{\Omega}_i \otimes \boldsymbol{\Phi}_i$ is the Kronecker product of $\boldsymbol{\Omega}_i$ and $\boldsymbol{\Phi}_i$. Note that $\boldsymbol{\Omega}_i$, $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Sigma}_i$ are i.i.d. for $i = 1, \dots, n$, and that $\boldsymbol{\Omega}_{i_1}$ and $\boldsymbol{\Phi}_{i_2}$ are independent for any i_1 and i_2. By simple calculations, we derive

$$\begin{aligned}\mathbf{a}^T \mathbf{B}\mathbf{B}^T \mathbf{a} &= \sum_{i=1}^n \big(\mathbf{B}_i^T \mathbf{a}\big)^2 = \sum_{i=1}^n \big[\mathrm{Tr}\big(\mathbf{A}^T \boldsymbol{\phi}(t_i)\mathbf{W}_i^T\big)\big]^2 \\ &= \sum_{i=1}^n \mathbf{W}_i^T \mathbf{A}^T \boldsymbol{\phi}(t_i)\boldsymbol{\phi}^T(t_i)\mathbf{A}\mathbf{W}_i = \sum_{i=1}^n \mathbf{a}^T (\boldsymbol{\Omega}_i \otimes \boldsymbol{\Phi}_i)\mathbf{a},\end{aligned}$$

which implies

$$\mathbf{B}^T \mathbf{B} = \sum_{i=1}^n \boldsymbol{\Omega}_i \otimes \boldsymbol{\Phi}_i. \tag{1.6}$$

Let

$$\widehat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{B}^T \mathbf{B} = n^{-1} \sum_{i=1}^n \boldsymbol{\Sigma}_i. \tag{1.7}$$

Then, due to the i.i.d. structure of the observations, one has

$$\begin{aligned}&\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{\Sigma}_1 = \boldsymbol{\Omega} \otimes \boldsymbol{\Phi} \qquad \text{with } \boldsymbol{\Omega} = \mathbb{E}\big(\mathbf{W}_1 \mathbf{W}_1^T\big) \quad \text{and} \\ &\boldsymbol{\Phi} = \mathbb{E}\big(\boldsymbol{\phi}(t_1)\boldsymbol{\phi}^T(t_1)\big).\end{aligned} \tag{1.8}$$

2. Assumptions and notation.

2.1. *Notation.* In what follows, we use bold script for matrices and vectors, for example, $\mathbf{A}$ or $\mathbf{a}$, and superscripts to denote elements of those matrices and

vectors, for example, $\mathbf{A}^{(i,j)}$ or $\mathbf{a}^{(j)}$. Below, we provide a brief summary of the notation used throughout this paper:

- For any vector $\mathbf{a} \in \mathbb{R}^p$, denote the standard l_1 and l_2 vector norms by $\|\mathbf{a}\|_1$ and $\|\mathbf{a}\|_2$, respectively. For vectors $\mathbf{a}, \mathbf{c} \in \mathbb{R}^p$, denote their scalar product by $\langle \mathbf{a}, \mathbf{c} \rangle$.
- For any function $q(t)$, $t \in [0,1]$, $\|q\|_2$ and $\langle \cdot, \cdot \rangle_2$ are, respectively, the norm and the scalar product in the space $L_2([0,1])$. Also, $\|q\|_\infty = \sup_{t\in[0,1]} |q(t)|$.
- For any vector function $\mathbf{q}(t) = (q_1(t), \dots, q_p(t))^T$, denote

$$\|\mathbf{q}(t)\|_2 = \left[\sum_{j=1}^{p} \|q_j\|_2^2 \right]^{1/2}.$$

- For any matrix $\mathbf{A}$, denote its spectral and Frobenius norms by $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_2$, respectively.
- Denote the $k \times k$ identity matrix by $\mathbb{I}_k$.
- For any numbers, a and b, denote $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$.
- In what follows, we use the symbol C for a generic positive constant, which is independent of n, p, s and l, and may take different values at different places.
- If $\mathbf{r} = (r_1, \dots, r_p)^T$ and $r_j' = r_j + 1/2 - 1/\nu_j$ for some $1 \le \nu_j < \infty$, denote $r_j^* = r_j \wedge r_j'$ and $r_{\min}^* = \min_j r_j^*$.
- Denote

$$\mathbf{t} = (t_1, \dots, t_n), \qquad \mathbb{W} = (\mathbf{W}_1, \dots, \mathbf{W}_n); \tag{2.1}$$

that is, $\mathbb{W}$ is the $p \times n$ matrix with columns $\mathbf{W}_i$, $i = 1, \dots, n$.

2.2. *Assumptions*. We impose the following assumptions on the varying coefficient model (1.3).

(A0). Only s out of p functions f_j are nonconstant and depend on the time variable t, s_0 functions are constant and independent of t and $(p - s - s_0)$ functions are identically equal to zero. We denote by $\mathcal{J}$ the set of indices corresponding to the nonconstant functions f_j.

(A1). Functions $(\phi_k(\cdot))_{k=0,\dots,\infty}$ form an orthonormal basis of $L_2([0,1])$, and are such that $\phi_0(t) = 1$ and, for any $t \in [0,1]$, any $l \ge 0$ and some $C_\phi < \infty$

$$\sum_{k=0}^{l} \phi_k^2(t) \le C_\phi^2 (l+1). \tag{2.2}$$

(A2). The probability density function $g(t)$ is bounded above and below $0 < g_1 \le g(t) \le g_2 < \infty$. Moreover, the eigenvalues of $\mathbb{E}(\boldsymbol{\phi}\boldsymbol{\phi}^T) = \boldsymbol{\Phi}$ are bounded from above and below

$$0 < \phi_{\min} = \lambda_{\min}(\boldsymbol{\Phi}) \le \lambda_{\max}(\boldsymbol{\Phi}) = \phi_{\max} < \infty.$$

Here, $\phi_{\min}$ and $\phi_{\max}$ are absolute constants independent of L.

(A3). Functions $f_j(t)$ have efficient representations in basis ϕ_l, in particular, for any $j = 1, \dots, p$, one has

$$\sum_{k=0}^{\infty} |a_{jk}|^{\nu_j} (k+1)^{\nu_j r_j'} \le (C_a)^{\nu_j}, \qquad r_j' = r_j + 1/2 - 1/\nu_j, \tag{2.3}$$

for some $C_a > 0$, $1 \le \nu_j < \infty$ and $r_j > \min(1/2, 1/\nu_j)$. In particular, if function $f_j(t)$ is constant or vanishes, then $r_j = \infty$. We denote vectors with elements ν_j and r_j, $j = 1, \dots, p$, by $\boldsymbol{\nu}$ and $\mathbf{r}$, respectively, and the set of indices of finite elements r_j by $\mathcal{J}$,

$$\mathcal{J} = \{j : r_j < \infty\}. \tag{2.4}$$

(A4). We suppose that the variables ξ_i, $i = 1, \dots, n$, are i.i.d., centered and *sub-Gaussian*. That is, $\mathbb{E}(\xi_i) = 0$, and there exists a constant K such that

$$\mathbb{E}\big[\exp(t\xi_i)\big] \le \exp\big(t^2 K^2/2\big)$$

for all $t > 0$. (See [33] for discussion of sub-Gaussian random variables.)

(A5). "Restricted isometry in expectation" condition. Let $\mathbf{W}_\Lambda$, $\Lambda \subset \{1, \dots, p\}$ be the sub-vector obtained by extracting the entries of $\mathbf{W}$ corresponding to indices in Λ, and let $\boldsymbol{\Omega}_\Lambda = \mathbb{E}(\mathbf{W}_\Lambda \mathbf{W}_\Lambda^T)$. We assume that there exist two positive constants $\omega_{\max}(\aleph)$ and $\omega_{\min}(\aleph)$ such that for all subsets Λ with cardinality $|\Lambda| \le \aleph$ and all $\mathbf{v} \in \mathbf{R}^{|\Lambda|}$ one has

$$\omega_{\min}(\aleph)\|\mathbf{v}\|_2^2 \le \mathbf{v}^T \boldsymbol{\Omega}_\Lambda \mathbf{v} \le \omega_{\max}(\aleph)\|\mathbf{v}\|_2^2. \tag{2.5}$$

Moreover, we suppose that $\mathbb{E}(\mathbf{W}^{(j)})^4 \le V$ for any $j = 1, \dots, p$, and that, for any $\mu \ge 1$ and for all subsets Λ with $|\Lambda| \le \aleph$, there exist positive constants U_μ and C_μ and a set $\mathcal{W}_\mu$ such that

$$\begin{aligned} &\mathbf{W}_\Lambda \in \mathcal{W}_\mu \Longrightarrow \big(\|\mathbf{W}_\Lambda\|_2 \le U_\mu\big) \cap \Big(\max_{j\in\Lambda} |\mathbf{W}_\Lambda^{(j)}| \le C_\mu\Big), \\ &\mathbb{P}(\mathbf{W}_\Lambda \in \mathcal{W}_\mu) \ \ge \ 1 - 2p^{-2\mu}. \end{aligned} \tag{2.6}$$

Here, $U_\mu = U_\mu(\aleph)$, $C_\mu = C_\mu(\aleph)$.

(A6). We assume that $(s + s_0)(1 + \log n) \le p$ and that there exists a numerical constant $C_\omega > 1$ such that

$$\log(n) \ge \frac{C_\omega \phi_{\max} \omega_{\max}((s+s_0)\log n)}{\phi_{\min}\omega_{\min}((s+s_0)(1+\log n))}. \tag{2.7}$$

We denote

$$\omega_{\max}^* = \omega_{\max}\big((s+s_0)\log n\big), \qquad \omega_{\min}^* = \omega_{\min}\big((s+s_0)(1+\log n)\big). \tag{2.8}$$

Note that $\omega_{\max}^*$ and $\omega_{\min}^*$ are functions of $((s+s_0)\log n)$ and $((s+s_0)(1+\log n))$, respectively.

2.3. *Discussion of assumptions.*

- Assumptions (A0) corresponds to the case when s of the covariates $f_j(t)$ are indeed functions of time, s_0 of them are time independent and $(p - s - s_0)$ are irrelevant.
- Assumption (A1) deals with the basis of $L_2([0, 1])$. There are many types of orthonormal bases satisfying those conditions.

  (a) *Fourier basis*. Choose $\phi_0(t) = 1$, $\phi_k(t) = 2\sin(2\pi kt)$ if $k > 1$ is odd, $\phi_k(t) = 2\cos(2\pi kt)$ if $k > 1$ is even. The basis functions are bounded and $C_\phi = 2$.

  (b) *Wavelet basis*. Consider a periodic wavelet basis on $[0, 1]$: $\psi_{h,i}(t) = 2^{h/2}\psi(2^h t - i)$ with $h = 0, 1, \ldots, i = 0, \ldots, 2^h - 1$. Set $\phi_0(t) = 1$ and $\phi_j(t) = \psi_{h,i}(t)$ with $j = 2^h + i + 1$. If $l = 2^J$, then condition (2.2) is satisfied with $C_\phi = \|\psi\|_\infty$. Observing that, for $2^J < l < 2^{J+1}$, we have $(l + 1) \geq (2^{J+1} + 1)/2$, and one can take $C_\phi = 2\|\psi\|_\infty$.
- Assumption (A2) that $\phi_{\min}$ and $\phi_{\max}$ are absolute constants independent of L is guaranteed by the fact that the sampling density g is bounded above and below. For example, if $g(t) = 1$, one has $\phi_{\min} = \phi_{\max} = 1$.
- Assumption (A3) describes sparsity of the vectors of coefficients of functions $f_j(t)$ in basis ϕ_l, $j = 1, \ldots, p$ and its smoothness. For example, if $\nu_j < 2$, the vector of coefficients a_{jl} of f_j is sparse. In the case when basis ϕ_l is formed by wavelets, condition (2.3) implies that f_j belongs to a Besov ball of radius C_a. If we chose Fourier bases and $\nu_j = 2$, then f_j belongs to a Sobolev ball of smoothness r_j and radius C_a. Note that assumption (A3) allows each nonconstant function f_j to have its own sparsity and smoothness patterns.
- Assumption (A4) that ξ_i are sub-Gaussian random variables means that their distribution is dominated by the distribution of a centered Gaussian random variable. This is a convenient and reasonably wide class. Classical examples of sub-Gaussian random variables are Gaussian, Bernoulli and all bounded random variables.
- Assumption (A5) is closely related to the restricted isometry (RI) conditions usually considered in the papers that employ LASSO technique or its versions; see, for example, [2]. However, usually the RI condition is imposed on the matrix of scalar products of the elements of a deterministic dictionary while we deal with a random dictionary and require this condition to hold only for the expectation of this matrix.

  Note that the upper bound in condition (2.5) is automatically satisfied with $\omega_{\max} = \|\mathbf{\Omega}\|$ where $\|\mathbf{\Omega}\|$ is the spectral norm of the matrix $\mathbb{E}(\mathbf{W}\mathbf{W}^T) = \mathbf{\Omega}$. If the smallest eigenvalue of $\mathbf{\Omega}$, $\lambda_{\min}(\mathbf{\Omega})$, is nonzero, then the lower bound in (2.5) is satisfied with $\omega_{\min} = \lambda_{\min}(\mathbf{\Omega})$. However, since the λ-restricted maximal eigenvalue $\omega_{\max}(\lambda)$ may be much smaller than the spectral norm of $\mathbf{\Omega}$ and $\omega_{\min}(\lambda)$ may be much larger then $\lambda_{\min}(\mathbf{\Omega})$, using those values will result in sharper

bounds for the error. Note that in the case when $\mathbf{W}$ has i.i.d. zero-mean entries W^j with $\mathbb{E}(W^{(j)})^2 = \nu^2$, we have $\omega_{\max} = \omega_{\min} = \nu^2$.

- Condition (2.7) is usual in the literature on the high-dimensional linear regression model; see, for example, Assumption 2 in [2]. For instance, if $\mathbf{W}$ has i.i.d. zero-mean entries W^j and $g(t) = 1$, this condition is satisfied for any $1 < C_\omega \le \log(n)$. Note that condition $(s+s_0)(1+\log n) \le p$ is slightly stronger than the usual condition $(s+s_0) \le 2p$. The additional $\log n$ factor corresponds to the block size.

3. Estimation strategy and nonasymptotic error bounds.

3.1. *Estimation strategy.* Formulation (1.4) implies that the varying coefficient model can be reduced to the linear regression model and one can apply one of the multitude of penalized optimization techniques which have been developed for the linear regression. In what follows, we apply a block LASSO penalties for the coefficients in order to account for both the constant and the vanishing functions f_j and also to take advantage of the sparsity of the functional coefficients in the chosen basis.

In particular, for each function f_j, $j = 1, \dots, p$, we divide its coefficients into $M+1$ different groups where group zero contains only coefficient a_{j0} for the constant function $\phi_0(t) = 1$ and M groups of size $d \approx \log n$ where $M = L/d$. We denote $\mathbf{a}_{j0} = a_{j0}$ and $\mathbf{a}_{jl} = (a_{j,d(l-1)+1}, \dots, a_{j,dl})^T$ the sub-vector of coefficients of function f_j in block l, $l = 1, \dots, M$. Let K_l be the subset of indices associated with $\mathbf{a}_{jl}$. We impose block norm on matrix $\mathbf{A}$ as follows:

$$\|\mathbf{A}\|_{\text{block}} = \sum_{j=1}^{p} \sum_{l=0}^{M} \|\mathbf{a}_{jl}\|_2. \tag{3.1}$$

Observe that $\|\mathbf{A}\|_{\text{block}}$ indeed satisfies the definition of a norm and is a sum of absolute values of coefficients a_{j0} of functions f_j and l_2 norms for each of the block vectors of coefficients $\mathbf{a}_{jl}$, $j = 1, \dots, p$, $l = 1, \dots, M$.

The penalty which we impose is related to both the ordinary and the group LASSO penalties which have been used by many authors. The difference, however, lies in the fact that the structure of the blocks is not motivated by naturally occurring groups (like, e.g., rows of the matrix $\mathbf{A}$) but rather our desire to exploit sparsity of functional coefficients a_{jl}. In particular, we construct an estimator $\widehat{\mathbf{A}}$ of $\mathbf{A}_0$ as a solution of the following convex optimization problem:

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A}} \left\{ n^{-1} \sum_{i=1}^{n} \left[Y_i - \operatorname{Tr}\left(\mathbf{A}^T \boldsymbol{\phi}(t_i) \mathbf{W}_i^T\right) \right]^2 + \delta \|\mathbf{A}\|_{\text{block}} \right\}, \tag{3.2}$$

where the value of δ will be defined later.

Note that with the tensor approach which we used in Section 1.2, optimization problem (3.2) can be re-written in terms of vector $\boldsymbol{\alpha} = \mathrm{Vec}(\mathbf{A})$ as

$$\hat{\mathbf{a}} = \arg\min_{\boldsymbol{\alpha}} \{n^{-1}\|\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}\|_2^2 + \delta\|\boldsymbol{\alpha}\|_{\mathrm{block}}\}, \tag{3.3}$$

where $\|\boldsymbol{\alpha}\|_{\mathrm{block}} = \|\mathbf{A}\|_{\mathrm{block}}$ is defined by the right-hand side (3.1) with vectors $\mathbf{a}_{jl}$ being sub-vectors of vector $\boldsymbol{\alpha}$. Subsequently, we construct an estimator $\hat{\mathbf{f}}(t) = (\hat{f}_1(t), \ldots, \hat{f}_p(t))^T$ of the vector function $\mathbf{f}(t)$ using

$$\hat{f}_j(t) = \sum_{k=0}^{L} \hat{a}_{jk}\phi_k(t), \qquad j = 1, \ldots, p. \tag{3.4}$$

In what follows, we derive the upper bounds for the risk of the estimator $\hat{\mathbf{a}}$ (or $\widehat{\mathbf{A}}$) and suggest a value of parameter δ, which allows us to attain those bounds. However, in order to obtain a benchmark of how well the procedure is performing, we determine the lower bounds for the risk of any estimator $\widehat{\mathbf{A}}$.

REMARK 1. Assumption that $K = L/d$ is an integer is not essential. Indeed, we can replace the number of groups K by the largest integer below or equal to L/d and then adjust group sizes to be d or $d+1$ where $d = [\log n]$, the largest integer not exceeding $\log n$.

3.2. *Lower bounds for the risk.* In this section we will obtain the lower bounds on the estimation risk. We consider a class $\mathcal{F} = \mathcal{F}_{s_0,s,\boldsymbol{\nu},\mathbf{r}}(C_a)$ of vector functions $\mathbf{f}(t)$ such that s of their components are nonconstant with coefficients satisfying condition (2.3) in (A3), s_0 of the components are constant and $(p - s - s_0)$ components are identically equal to zero. We construct the lower bound for the minimax quadratic risk of any estimator $\tilde{\mathbf{f}}$ of the vector function $\mathbf{f} \in \mathcal{F}_{s_0,s,\boldsymbol{\nu},\mathbf{r}}(C_a)$. Assume that, conditionally on $\mathbf{W}_i$ and t_i, variables ξ_i are Gaussian $\mathcal{N}(0, \sigma^2)$. Let $\omega^*_{\max}$ be given by formula (2.8). Denote $r_{\max} = \max\{r_j : j \in \mathcal{J}\}$ and

$$n_{\mathrm{low}} = \frac{2\sigma^2\kappa}{C_a^2\omega^*_{\max}\phi_{\max}} \max\left\{\frac{4\log(p/s_0)}{5}, \frac{8\log(p/s)}{5}, \left(\frac{6}{s}\right)^{2r_{\max}+1}\right\}, \tag{3.5}$$

$$\begin{aligned}\Delta_{\mathrm{lower}}(s_0, s, n, \mathbf{r}) = \max\Bigg\{&\frac{\kappa\sigma^2[s\log(p/s) + s_0\log(p/s_0)]}{5n\omega^*_{\max}\phi_{\max}},\\ &\frac{1}{8}\sum_{j\in\mathcal{J}} C_a^{2/(2r_j+1)}\left(\frac{\sigma^2\kappa}{n\omega^*_{\max}\phi_{\max}}\right)^{2r_j/(2r_j+1)}\Bigg\}.\end{aligned} \tag{3.6}$$

Then the following statement holds.

THEOREM 1. *Let $s \geq 1$ and $s_0 \geq 3$. Consider observations Y_i in model* (1.3) *with $\mathbf{W}_i$, $i = 1, \ldots, n$ and t satisfying assumptions* (A5) *and* (A2), *respectively.*

Assume that $n \geq n_{\mathrm{low}}$. *Then, for any* $\kappa < 1/8$ *and any estimator* $\tilde{\mathbf{f}}$ *of* $\mathbf{f}$, *one has*

$$\inf_{\tilde{\mathbf{f}}} \sup_{\mathbf{f}\in\mathcal{F}} \mathbb{P}\big(\|\tilde{\mathbf{f}} - \mathbf{f}\|_2^2 \geq \Delta_{\mathrm{lower}}(s_0, s, n, r)\big) \geq \frac{\sqrt{2}}{1+\sqrt{2}}\Big(1 - 2\kappa - \sqrt{\frac{2\kappa}{\log 2}}\Big). \tag{3.7}$$

Note that condition $s_0 \geq 3$ is not essential since, for $s_0 < 3$, the first term in (3.6) is of parametric order. Condition $n \geq n_{\mathrm{low}}$ is a purely technical condition which is satisfied for the collection of n's for which upper bounds are derived. Observe also that inequality (3.7) immediately implies that

$$\inf_{\tilde{\mathbf{f}}} \sup_{\mathbf{f}\in\mathcal{F}} \mathbb{E}\|\tilde{\mathbf{f}} - \mathbf{f}\|_2^2 \geq \Delta_{\mathrm{lower}}(s_0, s, n, r)\Big[\frac{\sqrt{2}}{1+\sqrt{2}}\Big(1 - 2\kappa - \sqrt{\frac{2\kappa}{\log 2}}\Big)\Big]. \tag{3.8}$$

3.3. *Adaptive estimation and upper bounds for the risk.* In this section we derive an upper bound for the risk of the estimator (3.2). For this purpose, first, we shall show that, with high probability, the ratio between the restricted eigenvalues of matrices $\widehat{\mathbf{\Sigma}}$ defined in (1.7) and $\mathbf{\Sigma} = \mathbb{E}\widehat{\mathbf{\Sigma}}$ is bounded above and below. This is accomplished by the following lemma.

For any $\Lambda \subset \{1, \dots, p\}$, we denote by $(\mathbf{\Omega}_i)_\Lambda = (\mathbf{W}_i)_\Lambda((\mathbf{W}_i)_\Lambda)^T$, $\widehat{\mathbf{\Sigma}}_\Lambda = n^{-1}\sum_{i=1}^n (\mathbf{\Omega}_i)_\Lambda \otimes \mathbf{\Phi}_i$ and $\mathbf{\Sigma}_\Lambda = \mathbb{E}\widehat{\mathbf{\Sigma}}_\Lambda$. For some $0 < h < 1$ and $1 \leq \aleph \leq p$, we define

$$N(\aleph) = \frac{64\mu\aleph(L+1)\log(p+L)C_\phi^2 U_\mu^2(\aleph)\phi_{\max}\omega_{\max}(\aleph)}{h^2\phi_{\min}^2\omega_{\min}^2(\aleph)}. \tag{3.9}$$

LEMMA 1. *Let* $n \geq N(\aleph)$ *and* μ *in* (2.6) *be large enough, so that*

$$p^\mu \geq \max\Big\{\frac{\sqrt{2V}n}{8\mu\sqrt{\aleph}U_\mu^2(\aleph)\log(p+L)}, 2n\Big\}, \tag{3.10}$$

where V *is defined in assumption* (A5). *Then, for any* $\Lambda \subset \{1, \dots, p\}$,

$$\inf_{\Lambda\,:\,|\Lambda|\leq\aleph} \big[\mathbb{P}\big(\big\{\|\widehat{\mathbf{\Sigma}}_\Lambda - \mathbf{\Sigma}_\Lambda\| < h\phi_{\min}\omega_{\min}(\aleph)\big\} \cap \mathcal{W}_\mu^{\otimes n}\big)\big] \geq 1 - 2p^{-\mu}, \tag{3.11}$$

where $\mathcal{W}_\mu$ *is the set of points in* $\mathbb{R}^p$ *such that condition* (2.6) *holds and* $\mathcal{W}_\mu^{\otimes n}$ *is the direct product of* n *sets* $\mathcal{W}_\mu$.

Moreover, on the set $\mathcal{W}_\mu^{\otimes n}$, *with probability at least* $1 - 2p^{-\mu}$, *one has simultaneously*

$$\begin{aligned} &\inf_{\Lambda\,:\,|\Lambda|\leq\aleph} \big[\lambda_{\min}(\widehat{\mathbf{\Sigma}}_\Lambda)\big] \geq (1-h)\phi_{\min}\omega_{\min}(\aleph), \\ &\sup_{\Lambda\,:\,|\Lambda|\leq\aleph} \big[\lambda_{\max}(\widehat{\mathbf{\Sigma}}_\Lambda)\big] \leq (1+h)\phi_{\max}\omega_{\max}(\aleph). \end{aligned} \tag{3.12}$$

Lemma 1 ensures that the restricted lowest eigenvalue of the regression matrix $\widehat{\mathbf{\Sigma}}$ is within a constant factor of the respective eigenvalue of matrix $\mathbf{\Sigma}$. Since p may be large, this is not guaranteed by a large value of n (as it happens in the

asymptotic setup) and leads to additional conditions on the relationship between parameters L, p and n.

The following theorem gives an upper bound for the quadratic risk of the estimator (3.2). We set $U_\mu = U_\mu(s+s_0)$. Denote

$$r_j^* = r_j \wedge r_j', \qquad r^* = \min_j r_j^*, \tag{3.13}$$

and suppose that $r^* \geq r_0^* > (2\varsigma)^{-1}$ where $1/2 \leq \varsigma < 1$ and choose $L+1 = n^\varsigma$. Let

$$\hat{\delta} = (C_K K\sqrt{\mu}+1)\sqrt{\frac{(1+h)\phi_{\max}\omega_{\max}(1)\log p}{n}}, \tag{3.14}$$

where constant C_K only depends on the distribution of random variables ξ_i and is introduced in (5.26) in [21]. Define

$$\begin{aligned}
\mathbf{N}_1 &= \frac{64\mu(s+s_0)C_\phi^2 U_\mu^2(L+1)\phi_{\max}\omega_{\max}^*\log(p+L)}{h^2\phi_{\min}^2(\omega_{\min}^*)^2},\\
\mathbf{N}_2 &= \frac{U_\mu^2 C_\phi^2(L+1)\mu\log p}{g_2\omega_{\max}(s)}, \qquad (3.15)\\
\mathbf{N}_3 &= \bigl(3C_a^2 g_2 s\omega_{\max}(s)/\log p\bigr)^{1/(2r_0^*\varsigma-1)},
\end{aligned}$$

and set $\mathbf{N} = \max(\mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3)$.

THEOREM 2. *Suppose that μ in (2.6) be large enough, so that*

$$p^\mu \geq \max\biggl\{\frac{\sqrt{2V}n}{8\mu\sqrt{s+s_0}U_\mu^2\log(p+L)}, \frac{2L}{\log n}, 2n\biggr\}. \tag{3.16}$$

Let $\hat{\mathbf{a}}$ be an estimator of $\mathbf{a}$ obtained as a solution of optimization problem (3.3) with $\delta = 2\hat{\delta}$, and the vector function $\hat{\mathbf{f}}$ be recovered using (3.4). If $n \geq \mathbf{N}$, then one has

$$\mathbb{P}\bigl(\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2 \leq \Delta(s_0,s,n,\mathbf{r})\bigr) \geq 1-8p^{-\mu}, \tag{3.17}$$

where

$$\begin{aligned}
\Delta(s_0,s,n,\mathbf{r}) = \frac{C_a^2 s}{n} &+ \frac{C_B(1+h)\omega_{\max}^*\phi_{\max}}{(1-h)\omega_{\min}^*\phi_{\min}}\\
&\times\biggl[\frac{(C_K K^2\mu+1)(s_0+s)\log p}{n(1-h)\omega_{\min}^*\phi_{\min}}\\
&\quad\times\sum_{j\in\mathcal{J}} C_a^{2/(2r_j+1)}\biggl(\frac{C_K K^2\mu+1}{n(1-h)\omega_{\min}^*\phi_{\min}}\biggr)^{2r_j/(2r_j+1)}\\
&\qquad\times(\log n)^{(2-\nu_j)_+/(\nu_j(2r_j+1))}\biggl(\frac{\log p}{\log n}\biggr)^{2r_j/(2r_j+1)}\biggr].
\end{aligned}$$

Additionally, the following inequality holds:

$$\mathbb{P}\big(n^{-1}\|\mathbf{W}^T(\hat{\mathbf{f}}-\mathbf{f})\|_2^2 \le \Delta'(s_0,s,n,\mathbf{r})\big) \ge 1-8p^{-\mu}, \tag{3.18}$$

where

$$\begin{aligned}\Delta'(s_0,s,n,\mathbf{r}) = \frac{C_a^2 s}{n} + C_B(1+h)\omega^*_{\max}\phi_{\max} \\ \times\Bigg[\frac{(C_K K^2\mu+1)(s_0+s)\log p}{n(1-h)\omega^*_{\min}\phi_{\min}} \\ \times\sum_{j\in\mathcal{J}} C_a^{2/(2r_j+1)}\bigg(\frac{C_K K^2\mu+1}{n(1-h)\omega^*_{\min}\phi_{\min}}\bigg)^{2r_j/(2r_j+1)} \\ \times(\log n)^{(2-\nu_j)_+/(\nu_j(2r_j+1))}\bigg(\frac{\log p}{\log n}\bigg)^{2r_j/(2r_j+1)}\Bigg].\end{aligned}$$

Note that construction (3.3) of the estimator $\hat{\mathbf{a}}$ does not involve knowledge of unknown parameters $\mathbf{r}$ and $\boldsymbol{\nu}$ or matrix $\boldsymbol{\Sigma}$. Therefore, estimator $\hat{\mathbf{a}}$ is fully adaptive with respect to those parameters. Moreover, conclusions of Theorem 2 are derived without any asymptotic assumptions on n, p and L.

REMARK 2. The combination of parameters ς and r_0^* provides a trade-off between adaptivity to the smoothness, sparsity and number of observations required for carrying out adaptive estimation. This relationship is not surprising since a larger number of dictionary functions used in the representation of $\mathbf{f}$ implies that we need to estimate a larger number of coefficients. In order to keep the estimator optimal, one has to either reduce the number of parameters in the model by considering a subclass of smoother functions (larger r_0^*, smaller L) or increase the number of observations (larger $\mathbf{N}_3$).

Indeed, if ς is a constant close to one, then our estimator adapts to the smoothness $r^* \ge (2\varsigma)^{-1}$ which is close to $1/2$. However, this will imply that $\mathbf{N}_3 \ge (3C_a^2 g_2 s\omega_{\max}(s)/\log p)^\nu$ where ν is very large; hence, either $s/\log p$ has to be very small or the number of observations n has to be very large. For larger values of s and without increasing the number of observations too much, one should choose smaller values of ς. For example, if $\varsigma = 1/2$, then the estimator adapts to smoothness $r^* \ge r_0^* = 2$ provided $n \ge \mathbf{N}_3 = 3C_a^2 g_2 s\omega_{\max}(s)/\log p$. If $\varsigma = 3/4$, then the estimator adapts to smoothness $r^* \ge r_0^* = 1$ provided $n \ge \mathbf{N}_3 = (3C_a^2 g_2 s\omega_{\max}(s)/\log p)^2$.

In order to assess the optimality of estimator $\hat{\mathbf{a}}$, we consider the case of the Gaussian noise; that is, ξ_i are Gaussian $\mathcal{N}(0,\sigma^2)$ and $K=\sigma$. Observe that, under assumption (A2), the values of $\phi_{\min}$ and $\phi_{\max}$ are independent of n and p,

so that the only quantities in (3.17) which are not bounded from above and below by an absolute constant are σ, $\omega^*_{\min}$, $\omega^*_{\max}$, s and s_0. Hence, $\Delta(s_0, s, n, \mathbf{r}) \le C\Delta_{\text{upper}}(s_0, s, n, \mathbf{r})$ with

$$
\begin{aligned}
&\Delta_{\text{upper}}(s_0, s, n, \mathbf{r}) \\
&\quad = \frac{\omega^*_{\max}}{\omega^*_{\min}}\Bigg[\frac{\sigma^2(s_0+s)\log p}{n\omega^*_{\min}} \\
&\qquad + \sum_{j\in\mathcal{J}} C_a^{2/(2r_j+1)}\left(\frac{\sigma^2}{n\omega^*_{\min}}\right)^{2r_j/(2r_j+1)} \\
&\qquad\quad \times (\log n)^{(2-\nu_j)_+/(\nu_j(2r_j+1))}\left(\frac{\log p}{\log n}\right)^{2r_j/(2r_j+1)}\Bigg],
\end{aligned}
\tag{3.19}
$$

where C is an absolute constant independent of C_a, n, p, σ^2 and vectors $\boldsymbol{\nu}$ and $\mathbf{r}$.

Inequality (3.19) implies that for any values of the parameters, the ratio between the upper and the lower bound for risk (3.6) is bounded by $C\varpi(n, p)(\omega^*_{\max}/\omega^*_{\min})^2$ where

$$
\begin{aligned}
\varpi(n, p) &= \max\left(\frac{\log p}{\log(p/s_0)}, \frac{\log p}{\log(p/s)}\right) \\
&\quad + \max_{j\in\mathcal{J}}(\log n)^{(2-\nu_j)_+/(\nu_j(2r_j+1))}\left(\frac{\log p}{\log n}\right)^{2r_j/(2r_j+1)}.
\end{aligned}
\tag{3.20}
$$

Note that $\omega^*_{\max}/\omega^*_{\min}$ is bounded by the condition number of matrix $\mathbf{\Omega}_\Lambda$ with $|\Lambda| = (s + s_0)(1 + \log n)$. Hence, if matrix $\mathbf{\Omega}_\Lambda$ is well conditioned, so that $\omega^*_{\max}/\omega^*_{\min}$ is bounded by a constant, the estimator $\hat{\mathbf{f}}$ attains optimal convergence rates up to a $\varpi(n, p)$ factor.

Suppose now that all functions $f_j(t)$ are spatially homogeneous, that is, $\min_j \nu_j \ge 2$. Additionally, assume that s_0 and s are small enough, that is, $s_0 \le p^{\gamma_1}$ and $s \le \max(\exp[Cn^{1/(2r+1)}], p^{\gamma_2})$ for some $\gamma_1 < 1$, $\gamma_2 < 1$, $C > 0$, and n is large enough, that is, there exists a positive β such that $n^\beta \ge p$. Then it is easy to see that $\varpi(n, p)$ is uniformly bounded, and the estimator $\hat{\mathbf{f}}$ attains optimal convergence rates up to a constant factor. In particular, if all functions in assumption (A3) belong to the same space, then the following corollary is valid.

COROLLARY 1. *Let the conditions of Theorem 2 hold with $r_j = r$ and $\nu_j = \nu$, $j - 1, \dots, p$, and matrix $\mathbf{\Omega}$ be well conditioned; that is, $\omega^*_{\max}/\omega^*_{\min}$ is bounded by some absolute constant independent of n, p and K. Let there exist $\beta > 0$, $\gamma_1 < 1$, $\gamma_2 < 1$ and $C > 0$ such that $n^\beta \ge p$, $s_0 \le p^{\gamma_1}$ and $s \le \max(\exp[Cn^{1/(2r+1)}], p^{\gamma_2})$. Then*

$$
\frac{\Delta_{\text{upper}}(s_0, s, n, \mathbf{r})}{\Delta_{\text{lower}}(s_0, s, n, \mathbf{r})} \le \begin{cases} C(\log n)^{2(2-\nu)_+/(\nu(2r+1))}, & \text{if } 1 \le \nu < 2, \\ C, & \text{if } \nu \ge 2. \end{cases}
\tag{3.21}
$$

As an example, consider the case when one knows that all functions f_j are polynomials. In this case, of course, one should use polynomial basis. If the maximum degree of the polynomials M is known, then one is dealing with a parametric estimation problem and obtains parametric convergence rates with s_0 replaced by $s_0 + sM$ and the second term is absent in (3.19). If the degrees of these polynomials can grow indefinitely, then the class of functions is nonparametric with $\nu \geq 2$, and we can be sure that we obtain convergence rates optimal up to a constant provided there exist $\beta > 0$ and $\gamma < 1$ such that $n^\beta \geq p$ and $s_0 \leq p^\gamma$.

3.4. *Adaptive estimation with respect to the mean squared risk.* Theorem 2 derives upper bounds for the risk with high probability. Suppose that an upper bound on the norms of functions $\mathbf{f}_j$ is available due to physical or other considerations,

$$\max_{1\leq j\leq p} \|f_j\|_2^2 \leq C_f^2. \tag{3.22}$$

Then $\|\mathbf{a}\|^2 \leq pC_f^2$ and $\hat{\mathbf{a}}$ given by (3.3) can be replaced by the solution of the convex problem

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \{n^{-1}\|\mathbf{Y} - \mathbf{B}\mathbf{a}\|_2^2 + \delta\|\mathbf{a}\|_{\text{block}} \text{ s.t. } \|\mathbf{a}\|^2 \leq pC_f^2\} \tag{3.23}$$

with $\delta = \hat{\delta}$ where $\hat{\delta}$ is defined in (3.14), and estimators $\hat{f}_j$ of f_j, $j = 1, \ldots, p$, are constructed using formula (3.4). Choose μ in (2.6) large enough, so that

$$16nC_f^2 \leq p^{\mu-1}. \tag{3.24}$$

Then the following statement is valid.

THEOREM 3. *Under the assumptions of the Theorem 2, and for any μ satisfying condition (3.24), one has*

$$\mathbb{E}\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2 \leq C\Delta_{\text{upper}}(s_0, s, n, \mathbf{r}), \tag{3.25}$$

where C is an absolute constant independent of n, p and K.

4. Examples and discussion.

4.1. *Examples.* In this section we provide several examples when assumptions of the paper are satisfied. For simplicity, we assume that $g(t) = 1$, so that $\phi_{\min} = \phi_{\max} = 1$.

EXAMPLE 1 (Normally distributed dictionary). Let $\mathbf{W}_i$, $i = 1, \ldots, n$, be i.i.d. standard Gaussian vectors $N(\mathbf{0}, \mathbf{I}_p)$. Then $\mathbf{\Omega} = \mathbf{I}_p$, so that $\omega_{\min} = \omega_{\max} = 1$. Moreover, $\mathbf{W}^{(j)}$ are independent standard Gaussian variables, and $(\mathbf{W}_i)_\Lambda^T(\mathbf{W}_i)_\Lambda$ are independent chi-squared variables with $|\Lambda| = \aleph$ degrees of freedom. Using the inequality (see, e.g., [3], page 67)

$$\mathbb{P}\big(\chi_\aleph^2 \leq \aleph + 2\sqrt{\aleph x} + 2x\big) \geq 1 - e^{-x}, \qquad x > 0,$$

for any $\mu_1 \geq 0$, derive

$$\mathbb{P}\big((\mathbf{W}_1)_\Lambda^T(\mathbf{W}_1)_\Lambda \leq (\sqrt{\aleph} + \sqrt{2}\mu_1)^2\big) \geq 1 - \exp(-\mu_1^2).$$

Choose any $\mu > 0$, and set $\mu_1^2 = 2\mu \log(p)$. Then using a standard bound on the maximum of p Gaussian variables, one obtains that assumption (A5) holds with

$$C_\mu = \sqrt{2\log p}, \qquad U_\mu^2 = (\sqrt{\aleph} + 2\sqrt{\mu \log p})^2.$$

EXAMPLE 2 (Symmetric Bernoulli dictionary). Let $\mathbf{W}_i^{(j)}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, be independent symmetric Bernoulli variables

$$\mathbb{P}(\mathbf{W}_i^{(j)} = 1) = \mathbb{P}(\mathbf{W}_i^{(j)} = -1) = 1/2.$$

Then $\mathbf{\Omega} = \mathbf{I}_p$, $\omega_{\min} = \omega_{\max} = 1$ and, for any μ,

$$C_\mu = 1, \qquad U_\mu^2(\aleph) = \aleph.$$

In both cases, $\mathbf{N}$ in (3.15) is of the form $\mathbf{N} = C(s + s_0)^2(L + 1)\log(p)$. Under the conditions of Theorem 2, the upper bounds for the risk are of the form

$$\begin{aligned}\Delta_{\text{upper}}&(s_0, s, n, \mathbf{r}) \\ &= C\bigg[\frac{K^2 s_0 \log p}{n} + \sum_{j \in \mathcal{J}} \bigg(\frac{K^2}{n}\bigg)^{2r_j/(2r_j+1)} C_a^{2/(2r_j+1)} \\ &\qquad \times (\log n)^{((2-\nu_j)_+ - 2\nu_j r_j)/(\nu_j(2r_j+1))} (\log p)^{2r_j/(2r_j+1)}\bigg],\end{aligned}$$

where C is a numerical constant. Now, it follows from Corollary 1 that the block LASSO estimator is minimax optimal up to, at most, the logarithmic factor of p.

The two examples above illustrate the situation when estimator (3.4) attains nearly optimal convergence rates when $p > n$. This, however, is not always possible. Note that our analysis of the performance of estimator (3.4) relies on the fact that the eigenvalues of any sub-matrix $\widehat{\mathbf{\Sigma}}_\Lambda$ are close to those of matrix $\mathbf{\Sigma}_\Lambda$ (Lemma 1). The latter requires $n \geq \mathbf{N}$ where $\mathbf{N}$ depends on the nature of vectors $\mathbf{W}_i$. The next example shows that sometimes $n < p$ prevents Lemma 1 from being valid.

EXAMPLE 3 (Orthonormal dictionary). Let $\mathbf{W}_i$, $i = 1, \ldots, n$, be uniformly distributed on a set of canonical vectors $\mathbf{e}_k$, $k = 1, \ldots, p$. Then $\mathbf{\Omega} = \mathbf{I}_p/p$, so that $\omega_{\min} = \omega_{\max} = 1/p$. Moreover, $\|\mathbf{W}_1\|_2^2 = 1$ and $|\mathbf{W}^{(j)}| \leq 1$. Therefore, for any $\mu > 0$,

$$C_\mu = 1, \qquad U_\mu^2(\aleph) = 1.$$

In the case of the orthonormal dictionary, $\mathbf{N}$ in (3.15) is of the form $\mathbf{N} = C(s + s_0)(L + 1)p\log(p)$. Under conditions of Theorem 2, the upper bound for the risk is of the form

$$\begin{aligned}
&\Delta_{\mathrm{upper}}(s_0, s, n, \mathbf{r}) \\
&\quad = C\Bigg[\frac{K^2 p s_0 \log n}{n} \\
&\qquad + \sum_{j\in\mathcal{J}} \left(\frac{pK^2}{n}\right)^{2r_j/(2r_j+1)} C_a^{2/(2r_j+1)} \\
&\qquad\qquad \times (\log n)^{((2-\nu_j)_+ - 2\nu_j r_j)/(\nu_j(2r_j+1))} (\log p)^{2r_j/(2r_j+1)}\Bigg],
\end{aligned}$$

so $n \geq \mathbf{N}$ implies $n > CK^2 p(s_0 + s)\log n$ which also guarantees that the risk of the estimator is small. This, indeed, coincides with one's intuition since one would need to sample more than p vectors in order to ensure that each component of the vector has been sampled at least once.

4.2. *Discussion*. In the present paper, we provide a nonasymptotic minimax study of the sparse high-dimensional varying coefficient model. To the best of our knowledge, this has never been accomplished before. An important feature of our analysis is its flexibility: it distinguishes between vanishing, constant and time-varying covariates, and in addition, it allows the latter to be heterogeneous (i.e., to have different degrees of smoothness) and spatially inhomogeneous. In this sense, our setup is more flexible than the one usually uses in the context of additive or compound functional models; see, for example, [9] or [29].

Our estimator is obtained using a block LASSO approach which can be viewed as a version of the group LASSO, where groups do not occur naturally but are rather driven by the need to reduce the variance, as it is done, for example, in block thresholding. Since we used a tensor approach for derivation of the estimator, we believe that the results of the paper can be generalized to the case of the multivariate varying coefficient model studied in [38].

An important feature of our estimator is that it is fully adaptive to the unknown sparsity and the smoothness of the time-varying covariates. Indeed, application of the proposed block LASSO technique does not require the knowledge of the number of the nonzero components of $\mathbf{f}$. It only depends on the highest diagonal element of matrix $\mathbf{\Omega}$ which can be estimated with high precision, even when n is quite small due to Lemma 1.

Note that, even when p is larger than n, the vector function $\mathbf{f}$ is completely identifiable due to assumption (A5). We consider some examples of the dictionaries such that this assumption holds. The latter ensures identifiability of $\mathbf{f}$ provided $n \geq \mathbf{N}$ where $\mathbf{N}$ is specified for each type of the random dictionary and depends on

the sparsity level of $\mathbf{f}$. On the other hand, large values of p ensure great flexibility of the choice of $\mathbf{f}$, so one can hope to represent the data using only few components of it.

Finally, we want to comment on the situation when the requirement $n \geq \mathbf{N}$ is not met due to lack of sparsity or insufficient number of observations. In this case, $\mathbf{f}$ is not identifiable, and one cannot guarantee that $\hat{\mathbf{f}}$ is close to the true function $\mathbf{f}$. However, these kinds of situations occur in all types of high-dimensional problems.

5. Proofs.

5.1. *Proofs of the lower bounds for the risk.* In order to prove Theorem 1, we consider a set of test vector functions $\mathbf{f}_{\boldsymbol{\omega}}(t) = (f_{1,\boldsymbol{\omega}}, \ldots, f_{p,\boldsymbol{\omega}})^T$ indexed by binary sequences $\boldsymbol{\omega}$ with components

$$f_{k,\boldsymbol{\omega}}(t) = \omega_{k0} u_k + \sum_{l=l_{0k}}^{2l_{0k}-1} \omega_{kl} v_k \phi_l(t), \tag{5.1}$$

where $l_{0k} \geq 1$ and $\omega_{kl} \in \{0, 1\}$ for $l = l_{0k}, l_{0k}+1, \ldots, 2l_{0k}-1$, $k = 1, \ldots, p$. Let K_0, K_1 and K_2 be disjoint sets of indices such that

$$f_{k,\boldsymbol{\omega}}(t) = \omega_{k0} u \qquad \text{if } k \in K_0, \tag{5.2}$$

$$f_{k,\boldsymbol{\omega}}(t) = \omega_{k0} \tilde{u} + \tilde{v}\phi_1(t) \qquad \text{if } k \in K_1 \tag{5.3}$$

and

$$f_{k,\boldsymbol{\omega}}(t) = v \sum_{l=l_{0k}}^{2l_{0k}-1} \omega_{kl} \phi_l(t) \qquad \text{if } k \in K_2. \tag{5.4}$$

In order for assumption (2.3) to hold, one needs $u \leq C_a$, $\tilde{u} \leq C_a/2$, $\tilde{v} \leq C_a 2^{-(r_k+1/2)}$ and

$$v^{\nu_j} \sum_{l=l_{0k}}^{2l_{0k}-1} (l+1)^{\nu_j r_j'} \leq (C_a)^{\nu_j}, \qquad j \in \Upsilon. \tag{5.5}$$

By simple calculations, it is easy to verify that condition (5.5) is satisfied if we set

$$v = C_a (2l_{0k})^{-(r_k+1/2)}, \tag{5.6}$$

where the constancy of v implies that l_{0k} in (5.6) are different for different values of k.

Consider two binary sequences $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ and the corresponding test functions $\mathbf{f}(t) = \mathbf{f}_{\boldsymbol{\omega}}(t)$ and $\tilde{\mathbf{f}}(t) = \mathbf{f}_{\tilde{\boldsymbol{\omega}}}(t)$ indexed by those sequences. Then the total squared distance in $L_2([0, 1])$ between $\mathbf{f}_{\boldsymbol{\omega}}(t)$ and $\mathbf{f}_{\tilde{\boldsymbol{\omega}}}(t)$ is equal to

$$\begin{aligned} D^2 = {} & u^2 \sum_{k \in K_0} |\omega_{k0} - \tilde{\omega}_{k0}| + \tilde{u}^2 \sum_{k \in K_1} |\omega_{k0} - \tilde{\omega}_{k0}| \\ & + v^2 \sum_{k \in K_2} \sum_{l=l_{0k}}^{2l_{0k}-1} |\omega_{kl} - \tilde{\omega}_{kl}|. \end{aligned} \tag{5.7}$$

Let $P_{\mathbf{f}}$ and $P_{\tilde{\mathbf{f}}}$ be probability measures corresponding to test functions $\mathbf{f}$ and $\tilde{\mathbf{f}}$, respectively. We shall consider three cases. In case 1, s_0 functions are constants of the form (5.2), and the rest of the functions are equal to identical zero. In case 2, s functions are time-dependent of the form (5.3), and the rest of the functions are equal to identical zero. In case 3, s functions are time-dependent of the form (5.4), and the rest of the functions are equal to identical zero. In all three cases, $\mathbf{f}(t)$ and $\tilde{\mathbf{f}}(t)$ contain at most $(s+s_0)$ nonzero coordinates. Using the fact that conditionally on W_i and t_i, the variables ξ_i are Gaussian $\mathcal{N}(0,\sigma^2)$, we obtain that the Kullback–Leibler divergence $\mathcal{K}(P_{\mathbf{f}}, P_{\tilde{\mathbf{f}}})$ between $P_{\mathbf{f}}$ and $P_{\tilde{\mathbf{f}}}$ satisfies

$$\mathcal{K}(P_{\mathbf{f}}, P_{\tilde{\mathbf{f}}}) = \bigl(2\sigma^2\bigr)^{-1}\mathbb{E}\sum_{i=1}^{n}\bigl[Q_i(\mathbf{f}) - Q_i(\tilde{\mathbf{f}})\bigr]^2 = \bigl(2\sigma^2\bigr)^{-1} n\mathbb{E}\bigl[Q_1(\mathbf{f}) - Q_1(\tilde{\mathbf{f}})\bigr]^2,$$

where

$$Q_i(\mathbf{f}) = W_i^T\mathbf{f}(t_i),$$

and, due to conditions (A2) and (A5),

$$\begin{aligned}\mathbb{E}\bigl[Q_1(\mathbf{f}) - Q_1(\tilde{\mathbf{f}})\bigr]^2 &= \mathbb{E}\bigl((\mathbf{f}-\tilde{\mathbf{f}})^T(t_1)W_1W_1^T(\mathbf{f}-\tilde{\mathbf{f}})(t_1)\bigr)\\ &= \mathbb{E}\bigl((\mathbf{f}-\tilde{\mathbf{f}})^T(t_1)\boldsymbol{\Omega}(\mathbf{f}-\tilde{\mathbf{f}})(t_1)\bigr)\\ &\le \omega^*_{\max}\mathbb{E}\bigl(\|(\mathbf{f}-\tilde{\mathbf{f}})(t_1)\|_2^2\bigr) \le \omega^*_{\max}\phi_{\max}D^2,\end{aligned}$$

where $\omega^*_{\max}$ and D^2 are defined in (2.8) and (5.7), respectively.

In order to derive the lower bounds for the risk, we use Theorem 2.5 of Tsybakov [31] which implies that if a set Θ of cardinality $(M+1)$ contains sequences $\boldsymbol{\omega}_0,\dots,\boldsymbol{\omega}_M$ with $M\ge 2$ such that, for any $j=1,\dots,M$, one has $\|f_{\boldsymbol{\omega}_0} - f_{\boldsymbol{\omega}_j}\| \ge D > 0$, $P_{\boldsymbol{\omega}_j} \ll P_{\boldsymbol{\omega}_0}$ and $\mathcal{K}(P_{\mathbf{f}_j}, P_{\mathbf{f}_0}) \le \kappa\log M$ with $0<\kappa<1/8$, then

$$\inf_{\tilde{\boldsymbol{\omega}}}\sup_{f_{\boldsymbol{\omega}},\boldsymbol{\omega}\in\Theta}\mathbb{P}\bigl(\|\mathbf{f}_{\boldsymbol{\omega}} - \mathbf{f}_{\tilde{\boldsymbol{\omega}}}\|_2 \ge D/2\bigr) \ge \frac{\sqrt{M}}{1+\sqrt{M}}\biggl(1 - 2\kappa - \sqrt{\frac{2\kappa}{\log M}}\biggr). \tag{5.8}$$

Now, we consider three separate cases.

Case 1. Let s_0 functions be constant, of the form (5.2), and the rest of the functions be identically equal to zero. Use Lemma 4.10 of Massart and Picard [26] (with $\alpha = 3/4$, $\beta = 1/4$ and $\rho = 0.233 \ge 0.2$) to choose a set Θ of binary sequences of length p with exactly s_0 ones such that the distance between any two sequences is at least $s_0/2$ and $\log[\operatorname{card}(\Theta)] \ge 0.2s_0\log(p/s_0)$. Then $D^2 \ge u^2 s_0/2$ and inequality

$$\mathcal{K}(P_{\mathbf{f}_j}, P_{\mathbf{f}_0}) \le \bigl(2\sigma^2\bigr)^{-1} n\omega^*_{\max}\phi_{\max}u^2 s_0/2 \le 0.2\kappa s_0\log(p/s_0)$$

holds if $u^2 = 4\sigma^2\kappa\log(p/s_0)/(5n\omega^*_{\max}\phi_{\max})$. Then

$$D^2 = \bigl(2s_0\log(p/s_0)\sigma^2\kappa\bigr)/\bigl(5n\omega^*_{\max}\phi_{\max}\bigr)$$

and $u \le C_a$ provided $n \ge 8\sigma^2\kappa\log(p/s_0)/(5C_a^2\omega^*_{\max}\phi_{\max})$.

Case 2. Let s functions be time-dependent of the form (5.3) and the rest of the functions be identically equal to zero. Use Lemma 4.10 of Massart and Picard [26] (with $\alpha = 3/4$, $\beta = 1/4$ and $\rho = 0.233 \geq 0.2$) to choose a set Θ of binary sequences of length p with exactly s ones such that the distance between any two sequences is at least $s/2$ and $\log[\operatorname{card}(\Theta)] \geq 0.2s \log(p/(s_0 + s))$. Then $D^2 \geq \tilde{u}^2 s/2$, and inequality

$$\mathcal{K}(P_{\mathbf{f}_j}, P_{\mathbf{f}_0}) \leq \bigl(2\sigma^2\bigr)^{-1} n\omega^*_{\max}\phi_{\max}\tilde{u}^2 s/2 \leq 0.2\kappa s \log(p/s)$$

holds if $\tilde{u}^2 = 4\sigma^2\kappa \log(p/s)/(5n\omega^*_{\max}\phi_{\max})$. Then

$$D^2 = \bigl(2s \log(p/s)\sigma^2\kappa\bigr)/\bigl(5n\omega^*_{\max}\phi_{\max}\bigr)$$

and $\tilde{u} \leq C_a/2$ provided $n \geq 16\sigma^2\kappa \log(p/s)/(5C_a^2\omega^*_{\max}\phi_{\max})$.

Case 3. Let the first s functions be time-dependent of the form (5.4) and the rest of the functions be equal to identical zero. Then $u = 0$, $\tilde{v} = 0$, v is given by formula (5.6) and $K_2 = \{1, \ldots, s\}$. Let r_k, $k \in K_2$ coincide with the values of finite components of vector $\mathbf{r}$. Denote

$$\mathcal{L} = \sum_{k=1}^{s} l_{0k}.$$

Use the Varshamov–Gilbert lemma to choose a set Θ of ω with $\operatorname{card}(\Theta) \geq 2^{\mathcal{L}/8}$ and $D^2 \geq v^2\mathcal{L}/8$. Inequality $\mathcal{K}(P_{\mathbf{f}_j}, P_{\mathbf{f}_0}) \leq \kappa\mathcal{L}/8$ holds if

$$v^2 \leq \bigl(\sigma^2\kappa\bigr)/\bigl(4n\omega^*_{\max}\phi_{\max}\bigr),$$

which, together with (5.6) and (5.7) imply that

$$l_{0k} = \left\lfloor \frac{1}{2}\left(\frac{4C_a^2 n\omega^*_{\max}\phi_{\max}}{\sigma^2\kappa}\right)^{1/(2r_k+1)} \right\rfloor + 1,$$

$$D^2 \geq \frac{C_a^2}{16}\sum_{k=1}^{s}\left(\frac{4C_a^2 n\omega^*_{\max}\phi_{\max}}{\sigma^2\kappa}\right)^{-2r_k/(2r_k+1)},$$

where $\lfloor x \rfloor$ denotes the integer part of x. Condition $\mathcal{L} \geq 3$ is satisfied for any $s \geq 1$ provided $n \geq n_{\text{low}}$.

5.2. *Proofs of the upper bounds for the risk.*

PROOF OF THEOREM 2. For any $\boldsymbol{\alpha} \in \mathbb{R}^{p(L+1)}$, one has

$$n^{-1}\|\mathbf{B}\hat{\boldsymbol{\alpha}} - \mathbf{Y}\|_2^2 + \delta\|\hat{\boldsymbol{\alpha}}\|_{\text{block}} \leq n^{-1}\|\mathbf{B}\boldsymbol{\alpha} - \mathbf{Y}\|_2^2 + \delta\|\boldsymbol{\alpha}\|_{\text{block}}.$$

Consider a set $\mathcal{F}_1 \subseteq \mathcal{W}_\mu^{\otimes n}$ such that (3.12) holds for any $\mathbf{t}$ and $\mathbb{W} \in \mathcal{F}_1$, and a set $\mathcal{F}_2 \subseteq \mathcal{W}_\mu^{\otimes n}$ such that (5.35) in [21] hold. Let $\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2$. Inequality (5.25) of Lemma 2 in [21] implies that, on the event $\mathcal{F}$,

$$\frac{2|\langle \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}, \mathbf{B}^T\mathbf{b}\rangle|}{n} \leq \hat{\delta}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\text{block}}.$$

Consider a set Ξ of values of the vector $\boldsymbol{\xi}$ such that

$$2n^{-1}|(\hat{\mathbf{a}}-\boldsymbol{\alpha},\mathbf{B}^T\boldsymbol{\xi})| \le \hat{\delta}\|\hat{\mathbf{a}}-\boldsymbol{\alpha}\|_{\text{block}} \qquad \text{for } \boldsymbol{\xi}\in\Xi. \tag{5.9}$$

By (5.24) in [21], one has $\mathbb{P}(\Xi)\ge 1-5p^{-\mu}$. Using (1.4), for $\boldsymbol{\xi}\in\Xi$, any $\mathbf{t}$ and $\mathbb{W}\in\mathcal{F}$, one obtains

$$\begin{aligned}\frac{\|\mathbf{B}(\hat{\mathbf{a}}-\mathbf{a})\|_2^2}{n} \le \frac{\|\mathbf{B}(\boldsymbol{\alpha}-\mathbf{a})\|_2^2}{n} - 2\hat{\delta}\|\hat{\mathbf{a}}\|_{\text{block}} \\ + 2\hat{\delta}\|\boldsymbol{\alpha}\|_{\text{block}} + \hat{\delta}\|\hat{\mathbf{a}}-\boldsymbol{\alpha}\|_{\text{block}}.\end{aligned} \tag{5.10}$$

Since $\boldsymbol{\alpha}$ is an arbitrary vector, setting $\boldsymbol{\alpha}=\mathbf{a}$ in (5.10) yields

$$\hat{\delta}\|\hat{\mathbf{a}}\|_{\text{block}} - \hat{\delta}\|\mathbf{a}\|_{\text{block}} \le 0.5\hat{\delta}\|\hat{\mathbf{a}}-\mathbf{a}\|_{\text{block}}.$$

Let the set J_0 contain the indices of nonzero blocks of $\mathbf{a}_0$,

$$J_0 = \{(j,l) : \|\mathbf{a}_{jl}\|_2 \ne 0\}. \tag{5.11}$$

Then the last inequality implies

$$\sum_{(i,j)\in J_0^C} \|(\mathbf{a}-\hat{\mathbf{a}})_{ij}\| \le 3\sum_{(i,j)\in J_0} \|(\mathbf{a}-\hat{\mathbf{a}})_{ij}\|. \tag{5.12}$$

From Lemma 1 it follows that

$$\lambda_{\min}(\widehat{\Sigma}) \ge (1-h)\phi_{\min}\omega^*_{\min} \quad \text{and} \quad \lambda_{\max}(\widehat{\Sigma}) \le (1+h)\phi_{\max}\omega^*_{\max},$$

where $\omega^*_{\min}$ and $\omega^*_{\max}$ are defined in (2.8). For $1\le j\le p$, consider the sets

$$G_{00} = \{j : 1\le j\le p, \alpha_{j0}=0\}, \qquad G_{01} = \{j : 1\le j\le p, \alpha_{j0}\ne 0\},$$
$$G_{j0} = \{l : 1\le l\le M, \|\boldsymbol{\alpha}_{jl}\|_2 = 0\}, \qquad G_{j1} = \{l : 1\le l\le M, \|\boldsymbol{\alpha}_{jl}\|_2 \ne 0\}.$$

We choose $\alpha_{j0}=a_{j0}$ if $a_{j0}\ne 0$ and $\alpha_{j0}=0$ otherwise. Let the sets G_{j1} be so that $l\in G_{j1}$ if and only if

$$\|\mathbf{a}_{jl}\|_2^2 > \varepsilon_0 = \frac{8^2(C_K K^2\mu+1)\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}.$$

We set $\boldsymbol{\alpha}_{jl}=\mathbf{a}_{jl}$ if $j\in\mathcal{J}$ and $l\in G_{j1}$ and $\boldsymbol{\alpha}_{jl}=\mathbf{0}$ otherwise where $\mathcal{J}$ is the set of indices corresponding to nonconstant functions f_j.

With $\delta=2\hat{\delta}$, inequality (5.12) and Lemma 3 in [21] guarantee that

$$\frac{\|\mathbf{B}(\hat{\mathbf{a}}-\mathbf{a})\|_2^2}{n} \ge C_B\lambda_{\min}(\widehat{\Sigma})\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2. \tag{5.13}$$

On the other hand, Lemma 1 and the definition of $\boldsymbol{\alpha}$ imply that

$$\frac{\|\mathbf{B}(\boldsymbol{\alpha}-\mathbf{a})\|_2^2}{n} \le \lambda_{\max}(\widehat{\Sigma})\|\boldsymbol{\alpha}-\mathbf{a}\|_2^2. \tag{5.14}$$

Then, using (5.13) and (5.14), we rewrite inequality (5.10) as

$$C_B\lambda_{\min}(\widehat{\Sigma})\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2 \le \lambda_{\max}(\widehat{\Sigma})\|(\boldsymbol{\alpha}-\mathbf{a})\|_2^2 + 4\hat{\delta}\sum_{j\in G_{01}}|\hat{a}_{j0}-a_{j0}| + 4\hat{\delta}\sum_{j\in\mathcal{J}}\sum_{l\in G_{j1}}\|\hat{\mathbf{a}}_{jl}-\mathbf{a}_{jl}\|_2.$$

Using inequality $2x_1x_2 \le x_1^2+x_2^2$ for any x_1, x_2, we derive

$$4\hat{\delta}\sum_{j\in\mathcal{J}}\sum_{l\in G_{j1}}\|\hat{\mathbf{a}}_{jl}-\mathbf{a}_{jl}\|_2 \le \frac{C_B\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}{2}\sum_{j\in\mathcal{J}}\sum_{l\in G_{j1}}\|\hat{\mathbf{a}}_{jl}-\mathbf{a}_{jl}\|_2^2 + \sum_{j\in\mathcal{J}}\frac{8\hat{\delta}^2\operatorname{card}(G_{j1})}{C_B\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})},$$

and similar inequality applies to the first sum in (5.15). By subtracting $0.5C_B\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2$ from both sides of (5.15) and plugging in the values of $\hat{\delta}$ and $\boldsymbol{\alpha}$, derive

$$\|(\hat{\mathbf{a}}-\mathbf{a})\|_2^2 \le \frac{C_B\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})}{\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\Bigg[\sum_{j=1}^{p}\sum_{l\in G_{j0}}\|\mathbf{a}_{jl}\|_2^2 + \frac{8^2(C_KK^2\mu+1)(s_0+s)\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})} + \sum_{j\in\mathcal{J}}\frac{8^2(C_KK^2\mu+1)\operatorname{card}(G_{j1})\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\Bigg]. \tag{5.15}$$

Observe that for any $1 \le j \le p$, one has

$$\sum_{l\in G_{j0}}\|\mathbf{a}_{jl}\|_2^2 + \frac{8^2(C_KK^2\mu+1)\operatorname{card}(G_1j)\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})} \le \sum_{l=1}^{M}\min\Bigg(\|\mathbf{a}_{jl}\|_2^2, \frac{8^2(C_KK^2\mu+1)\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\Bigg).$$

Application of inequality (5.31) in [21] with $\varepsilon = \frac{8^2(C_KK^2\mu+1)\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})\log n}$ (see Lemma 4 in [21]) yields

$$\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2 \le \frac{C_B\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})}{\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})} \times \Bigg[\frac{(C_KK^2\mu+1)(s_0+s)\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})} \times \sum_{j\in\mathcal{J}}C_a^{2/(2r_j+1)}\Bigg(\frac{C_KK^2\mu+1}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\Bigg)^{2r_j/(2r_j+1)} \times (\log n)^{((2-\nu_j)_+-2\nu_jr_j)/(\nu_j(2r_j+1))}(\log p)^{2r_j/(2r_j+1)}\Bigg]. \tag{5.16}$$

Using (5.33) in [21] and $L+1=n^{\varsigma}$, we obtain

$$\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2 \le \|\hat{\mathbf{a}}-\mathbf{a}\|_2^2 + C_a^2 s n^{-2r^*\varsigma}.$$

Condition $r^* \ge (2\varsigma)^{-1}$ yields

$$\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2 \le \|\hat{\mathbf{a}}-\mathbf{a}\|_2^2 + \frac{C_a^2 s}{n}.$$

Now, (5.16) together with Lemmas 1, 2 and 4 imply

$$\mathbb{P}\big(\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2 \le \Delta(s_0,s,n,\mathbf{r})\big) \ge 1-8p^{-\mu},$$

which proves upper bound (3.17) on the estimation error.

In order to prove upper bound (3.18) in the same way as (3.17), we derive the inequality

$$\begin{aligned}
&n^{-1}\|\mathbf{B}(\hat{\mathbf{a}}-\mathbf{a})\|_2^2 \\
&\quad\le C_B\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})\Bigg[\frac{(C_K K^2\mu+1)(s_0+s)\log p}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})} \\
&\qquad\times \sum_{j\in\mathcal{J}} C_a^{2/(2r_j+1)}\bigg(\frac{C_K K^2\mu+1}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\bigg)^{2r_j/(2r_j+1)} \\
&\qquad\times (\log n)^{((2-\nu_j)_+-2\nu_j r_j)/(\nu_j(2r_j+1))}(\log p)^{2r_j/(2r_j+1)}\Bigg].
\end{aligned}$$

Now, upper bound (3.18) on the prediction risk follows from Lemma 5 in [21], condition $L+1=n^{\varsigma}$ and $n^{-1}\|\mathbf{W}^T(\hat{\mathbf{f}}-\mathbf{f})\|_2^2 \le 2n^{-1}\|\mathbf{B}(\hat{\mathbf{a}}-\mathbf{a})\|_2^2 + 2n^{-1}\|\mathbf{b}\|_2^2$. □

PROOF OF THEOREM 3. Let sets $\tilde{\mathcal{F}}$ be such that (5.16) holds. Then $P(\tilde{\mathcal{F}}) \ge 1-8p^{-\mu}$ and (5.16) yields

$$\begin{aligned}
\mathbb{E}\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2 &\le \mathbb{E}\big[\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2\mathbb{I}(\tilde{\mathcal{F}})\big] + \mathbb{E}\big[\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2\mathbb{I}(\tilde{\mathcal{F}}^C)\big] \\
&\le C\frac{\omega^*_{\max}}{\omega^*_{\min}}\Bigg[\frac{K^2 s_0\log n}{n\omega^*_{\min}} \\
&\qquad + \sum_{j\in\mathcal{J}} C_a^{2/(2r_j+1)}\bigg(\frac{K^2}{n\omega^*_{\min}}\bigg)^{2r_j/(2r_j+1)} \\
&\qquad\times (\log n)^{((2-\nu_j)_+-2\nu_j r_j)/(\nu_j(2r_j+1))}(\log p)^{2r_j/(2r_j+1)}\Bigg] \\
&\quad + 16C_f^2 p^{1-\mu} \\
&\le C\Delta_{\text{upper}}(s_0,s,n,r),
\end{aligned}$$

due to (3.24). □

5.3. *Proof of Lemma 1.* In order to simplify the notation, we set $\omega_{\max}(\aleph) = \omega_{\max}$ and $U_\mu = U_\mu(\aleph)$. Let $\mathcal{W}_\mu$ be the set of points described in condition (2.6) of assumption (A5). Denote the direct product of n sets $\mathcal{W}_\mu$ by $\mathcal{W}_\mu^{\otimes n}$. Then

$$\mathbb{P}\big(\mathcal{W}_\mu^{\otimes n}\big) \geq 1 - 2np^{-2\mu}.$$

Consider random matrices

$$\mathbf{Z}_i = (\mathbf{\Sigma}_i)_\Lambda - \mathbf{\Sigma}_\Lambda = (\mathbf{\Omega}_i)_\Lambda \otimes \mathbf{\Phi}_i - \mathbf{\Omega}_\Lambda \otimes \mathbf{\Phi},$$
$$\boldsymbol{\zeta}_i = (\mathbf{\Sigma}_i)_\Lambda \mathbb{I}(\mathcal{W}_\mu) - \mathbb{E}\big((\mathbf{\Sigma}_i)_\Lambda \mathbb{I}(\mathcal{W}_\mu)\big).$$

Then $\boldsymbol{\zeta}_i$ are i.i.d. with $\mathbb{E}\boldsymbol{\zeta}_i = 0$. We apply the matrix version of Bernstein's inequality, given in Tropp [30]:

PROPOSITION 1 (Theorem 1.6, Tropp [30]). *Let $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n$ be independent random matrices in $\mathbb{R}^{m_1 \times m_2}$ such that $\mathbb{E}(\boldsymbol{\zeta}_i) = 0$. Define*

$$\sigma_\zeta = \max\left\{ \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\boldsymbol{\zeta}_i \boldsymbol{\zeta}_i^T) \right\|^{1/2}, \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\boldsymbol{\zeta}_i^T \boldsymbol{\zeta}_i) \right\|^{1/2} \right\},$$

and suppose that $\|\boldsymbol{\zeta}_i\| \leq T$ for some $T > 0$. Then, for all $t > 0$, with probability at least $1 - e^{-t}$, one has

$$\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\zeta}_i \right\| \leq 2 \max\left\{ \sigma_\zeta \sqrt{\frac{t + \log(d)}{n}}, T \frac{t + \log(d)}{n} \right\}, \tag{5.17}$$

where $d = m_1 + m_2$.

In order to find σ_ζ, note that

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\boldsymbol{\zeta}_i \boldsymbol{\zeta}_i^T) \right\| \\
&\quad = \big\| \mathbb{E}(\boldsymbol{\zeta}_1 \boldsymbol{\zeta}_1^T) \big\| \\
&\quad \leq \big\| \mathbb{E}\big((\mathbf{\Sigma}_1)_\Lambda (\mathbf{\Sigma}_1)_\Lambda^T \mathbb{I}(\mathcal{W}_\mu)\big) \big\| + \big\| \mathbb{E}\big((\mathbf{\Sigma}_1)_\Lambda \mathbb{I}(\mathcal{W}_\mu)\big) \big\| \big\| \mathbb{E}\big((\mathbf{\Sigma}_1)_\Lambda^T \mathbb{I}(\mathcal{W}_\mu)\big) \big\| \\
&\quad = \big\| \mathbb{E}\big[\big((\mathbf{\Omega}_1)_\Lambda \otimes \mathbf{\Phi}_1\big)\big((\mathbf{\Omega}_1)_\Lambda \otimes \mathbf{\Phi}_1\big) \mathbb{I}(\mathcal{W}_\mu) \big] \big\| + \big\| \mathbb{E}\big[\big((\mathbf{\Omega}_1)_\Lambda \otimes \mathbf{\Phi}_1\big) \mathbb{I}(\mathcal{W}_\mu) \big] \big\|^2 \\
&\quad = \big\| \mathbb{E}\big[\big((\mathbf{\Omega}_1)_\Lambda (\mathbf{\Omega}_1)_\Lambda\big) \otimes (\mathbf{\Phi}_1 \mathbf{\Phi}_1) \mathbb{I}(\mathcal{W}_\mu) \big] \big\| + \big\| \mathbb{E}\big((\mathbf{\Omega}_1)_\Lambda \mathbb{I}(\mathcal{W}_\mu)\big) \big\|^2 \big\| \mathbb{E}(\mathbf{\Phi}_1) \big\|^2 \\
&\quad \leq \big\| \mathbb{E}(\mathbf{\Phi}_1 \mathbf{\Phi}_1) \big\| \big\| \mathbb{E}\big[\big((\mathbf{\Omega}_1)_\Lambda (\mathbf{\Omega}_1)_\Lambda\big) \mathbb{I}(\mathcal{W}_\mu) \big] \big\| + \big\| \mathbb{E}\big((\mathbf{\Omega}_1)_\Lambda \mathbb{I}(\mathcal{W}_\mu)\big) \big\|^2 \big\| \mathbb{E}(\mathbf{\Phi}_1) \big\|^2,
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\boldsymbol{\zeta}_i^T \boldsymbol{\zeta}_i) \right\| &\leq \big\| \mathbb{E}(\mathbf{\Phi}_1 \mathbf{\Phi}_1) \big\| \big\| \mathbb{E}\big[\big((\mathbf{\Omega}_1)_\Lambda (\mathbf{\Omega}_1)_\Lambda\big) \mathbb{I}(\mathcal{W}_\mu) \big] \big\| \\
&\quad + \big\| \mathbb{E}\big((\mathbf{\Omega}_1)_\Lambda \mathbb{I}(\mathcal{W}_\mu)\big) \big\|^2 \big\| \mathbb{E}(\mathbf{\Phi}_1) \big\|^2.
\end{aligned}$$

Here,

$$\|\mathbb{E}(\mathbf{\Phi}_1\mathbf{\Phi}_1)\| \le \|C_\phi^2(L+1)\mathbf{\Phi}\| = C_\phi^2(L+1)\phi_{\max}$$

and

$$\begin{aligned}\|\mathbb{E}[((\mathbf{\Omega}_1)_\Lambda(\mathbf{\Omega}_1)_\Lambda)\mathbb{I}(\mathcal{W}_\mu)]\| &= \|\mathbb{E}[\mathbf{W}_\Lambda\mathbf{W}_\Lambda^T\mathbf{W}_\Lambda\mathbf{W}_\Lambda^T\mathbb{I}(\mathcal{W}_\mu)]\| \\ &\le U_\mu^2\|\mathbb{E}(\mathbf{W}_\Lambda\mathbf{W}_\Lambda^T)\| \\ &= U_\mu^2\|\mathbf{\Omega}_\Lambda\| = U_\mu^2\omega_{\max},\end{aligned}$$

so that

$$\sigma_\zeta^2 \le 2C_\phi^2U_\mu^2(L+1)\phi_{\max}\omega_{\max}. \tag{5.18}$$

Now, observe that, since matrix $\mathbb{E}((\mathbf{\Sigma}_i)_\Lambda\mathbb{I}(\mathcal{W}_\mu))$ is nonnegative definite and matrices $\mathbf{\Phi}_i$ and $(\mathbf{\Omega}_i)_\Lambda$ have rank one for any i, one has

$$\begin{aligned}T &= \sup\|\boldsymbol{\zeta}_1\| \le 2\sup\|(\mathbf{\Sigma}_1)_\Lambda\mathbb{I}(\mathcal{W}_\mu)\| \\ &= 2\sup\|(\mathbf{\Omega}_1)_\Lambda\mathbb{I}(\mathcal{W}_\mu)\|\|\mathbf{\Phi}_1\| \\ &\le 2C_\phi^2U_\mu^2(L+1).\end{aligned} \tag{5.19}$$

Apply Bernstein's inequality (5.17) with σ_ζ^2 and T given by formulas (5.18) and (5.19), respectively. Then we obtain for any $t>0$, with probability at least $1-e^{-t}$,

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{\zeta}_i\right\| \le 4\max\Bigg\{&\frac{C_\phi U_\mu\sqrt{(L+1)(t+\log(L\aleph)\phi_{\max}\omega_{\max})}}{\sqrt{n}}, \\ &\frac{C_\phi^2U_\mu^2(L+1)(t+\log(L\aleph))}{n}\Bigg\}.\end{aligned} \tag{5.20}$$

In order to apply inequality (5.20) to $\mathbf{Z}_i$, observe that $\mathbf{Z}_i - \boldsymbol{\zeta}_i = (\mathbf{\Sigma}_i)_\Lambda\mathbb{I}(\mathcal{W}_\mu^c) - \mathbb{E}(\mathbf{\Sigma}_i)_\Lambda\mathbb{I}(\mathcal{W}_\mu^c)$ and

$$\begin{aligned}\|\mathbb{E}((\mathbf{\Sigma}_i)_\Lambda\mathbb{I}(\mathcal{W}_\mu^c))\|^2 &= \|\mathbb{E}((\mathbf{\Omega}_i)_\Lambda\mathbb{I}(\mathcal{W}_\mu^c))\otimes\mathbb{E}(\mathbf{\Phi}_i)\|^2 \\ &\le \|\mathbb{E}((\mathbf{\Omega}_i)_\Lambda\mathbb{I}(\mathcal{W}_\mu^c))\|^2\|\mathbb{E}(\mathbf{\Phi}_i)\|^2 \\ &\le \big(\mathbb{E}\|(\mathbf{\Omega}_i)_\Lambda\mathbb{I}(\mathcal{W}_\mu^c)\|_2\big)^2\mathbb{E}\|\mathbf{\Phi}_i\|_2^2 \\ &\le 2C_\phi^4V\aleph p^{-2\mu}(L+1)^2,\end{aligned} \tag{5.21}$$

due to the fact that $\mathbb{E}(\mathbf{W}^{(j)})^4 \le V$ for any $j=1,\dots,p$.

Now, we use the union bound over all Λ such that $|\Lambda|\le\aleph$, the inequality $\binom{p}{\aleph} \le (\frac{ep}{\aleph})^\aleph$ and choose $t = 2\mu\aleph\log(\frac{ep}{\aleph})$. Combining (5.20) and (5.21), for all Λ such

that $|\Lambda| \le \aleph$, we derive

$$\begin{aligned}
&\inf_{\Lambda\,:\,|\Lambda|\le\aleph} \mathbb{P}\biggl(\biggl\{\biggl\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\biggr\| < z\biggr\} \cap \{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\}\biggr) \\
&\quad \ge \inf_{\Lambda\,:\,|\Lambda|\le\aleph} \mathbb{P}\biggl(\biggl\{\biggl\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\zeta}_i\biggr\| < z - \bigl\|\mathbb{E}\bigl((\boldsymbol{\Sigma}_1)_\Lambda \mathbb{I}\bigl(\mathcal{W}_\mu^c\bigr)\bigr)\bigr\|\biggr\} \cap \{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\}\biggr) \\
&\quad \ge 1 - \sup_{\Lambda\,:\,|\Lambda|\le\aleph} \mathbb{P}\biggl(\biggl\{\biggl\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\zeta}_i\biggr\| \ge z - C_\phi^2(L+1)p^{-\mu}\sqrt{2V\aleph}\biggr\} \\
&\qquad\qquad \cap \{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\}\biggr) \\
&\quad \ge 1 - (ep)^{-\mu} - 2np^{-2\mu}
\end{aligned}$$

for any z such that

$$\begin{aligned}
z \ge 8\max\biggl\{&\frac{C_\phi U_\mu \sqrt{\mu\aleph(L+1)\log(p+L)\phi_{\max}\omega_{\max}}}{\sqrt{n}}, \\
&\frac{\mu\aleph C_\phi^2 U_\mu^2 (L+1)\log(p+L)}{n}\biggr\} \\
&+ C_\phi^2(L+1)p^{-\mu}\sqrt{2V\aleph}.
\end{aligned} \tag{5.22}$$

Note that, under condition (3.10), one has

$$C_\phi^2(L+1)p^{-\mu}\sqrt{2V\aleph} \le 8\mu\aleph C_\phi^2 U_\mu^2 (L+1)\log(p+L)n^{-1}.$$

It is easy to check that, whenever $n \ge N(\aleph)$ where $N(\aleph)$ is defined in (3.9), condition (5.22) is satisfied with

$$\begin{aligned}
z &= \frac{8C_\phi U_\mu \sqrt{\mu\aleph(L+1)\phi_{\max}\omega_{\max}\log(p+L)}}{\sqrt{n}} + \frac{9\mu\aleph C_\phi^2 U_\mu^2 (L+1)\log(p+L)}{n} \\
&\le h\omega_{\min}\phi_{\min},
\end{aligned}$$

which, together with condition $p^\mu \ge 2n$, implies that

$$\inf_{\Lambda\,:\,|\Lambda|\le\lambda} \mathbb{P}\bigl(\|\widehat{\boldsymbol{\Sigma}}_\Lambda - \boldsymbol{\Sigma}_\Lambda\| \le h\omega_{\min}\phi_{\min}\bigr) \ge 1 - p^{-\mu} - p^{-\mu}. \tag{5.23}$$

In order to complete the proof, observe that $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_\Lambda) \ge \lambda_{\min}(\boldsymbol{\Sigma}_\Lambda) - \|\widehat{\boldsymbol{\Sigma}}_\Lambda - \boldsymbol{\Sigma}_\Lambda\|$ and $\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}_\Lambda) \le \lambda_{\max}(\boldsymbol{\Sigma}_\Lambda) + \|\widehat{\boldsymbol{\Sigma}}_\Lambda - \boldsymbol{\Sigma}_\Lambda\|$.

Acknowledgments. The authors want to thank Alexandre Tsybakov for extremely interesting suggestions and discussions. We also want to thank the referees whose valuable remarks and suggestions helped us to improve this paper.

SUPPLEMENTARY MATERIAL

Supplement to "Sparse high-dimensional varying coefficient model: Non-asymptotic minimax study" (DOI: 10.1214/15-AOS1309SUPP; .pdf). The supplemental material contains omitted proofs.

REFERENCES

[1] BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

[2] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[3] BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064

[4] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[5] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[6] CHESNEAU, CH. and HEBIRI, M. (2008). Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.* **17** 317–326. MR2483460

[7] CHIANG, C.-T., RICE, J. A. and WU, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* **96** 605–619. MR1946428

[8] CLEVELAND, W. S., GROSSE, E. and SHYU, W. M. (1991). Local regression models. In *Statistical Models in S* (J. M. Chambers and T. J. Hastie, eds.) 309–376. Wadsworth, Belmont, CA.

[9] DALALYAN, A., INGSTER, Y. and TSYBAKOV, A. B. (2014). Statistical inference in compound functional models. *Probab. Theory Related Fields* **158** 513–532. MR3176357

[10] FAN, J., MA, Y. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284. MR3265696

[11] FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

[12] FAN, J. and ZHANG, W. (2008). Statistical methods with varying coefficient models. *Stat. Interface* **1** 179–195. MR2425354

[13] HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

[14] HOOVER, D. R., RICE, J. A., WU, C. O. and YANG, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85** 809–822. MR1666699

[15] HUANG, J. Z. and SHEN, H. (2004). Functional coefficient regression models for nonlinear time series: A polynomial spline approach. *Scand. J. Stat.* **31** 515–534. MR2101537

[16] HUANG, J. Z., WU, C. O. and ZHOU, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128. MR1888349

[17] HUANG, J. Z., WU, C. O. and ZHOU, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972

[18] KAI, B., LI, R. and ZOU, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Ann. Statist.* **39** 305–332. MR2797848

[19] KAUERMANN, G. and TUTZ, G. (1999). On model diagnostics using varying coefficient models. *Biometrika* **86** 119–128. MR1688076

[20] KLOPP, O. and PENSKY, M. (2013). Nonasymptotic approach to varying coefficient model. *Electron. J. Stat.* **7** 454–479. MR3020429

[21] KLOPP, O. and PENSKY, M. (2015). Supplement to "Sparse high-dimensional varying coefficient model: nonasymptotic minimax study." DOI:10.1214/15-AOS1309SUPP.

[22] LEE, Y. K., MAMMEN, E. and PARK, B. U. (2012). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048

[23] LIAN, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statist. Sinica* **22** 1563–1588. MR3027099

[24] LIAN, H. and MA, S. (2013). Reduced-rank regression in sparse multivariate varying-coefficient models with high-dimensional covariates. Preprint. Available at arXiv:1309.6058v1.

[25] LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

[26] MASSART, P. and PICARD, J. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin.

[27] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

[28] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

[29] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427. MR2913704

[30] TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[31] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. MR2724359

[32] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

[33] VERSHYNIN, R. (2012). Introduction to the nonasymptotic analysis of random matrices. In *Compressed Sensing* (Y. Eldar and G. Kutyniok, eds.) 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[34] WANG, L., KAI, B. and LI, R. (2009). Local rank inference for varying coefficient models. *J. Amer. Statist. Assoc.* **104** 1631–1645. MR2597005

[35] WEI, F., HUANG, J. and LI, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica* **21** 1515–1540. MR2895107

[36] WU, C. O., CHIANG, C.-T. and HOOVER, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* **93** 1388–1402. MR1666635

[37] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[38] ZHU, H., LI, R. and KONG, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Statist.* **40** 2634–2666. MR3097615

92001 CREST AND MODAL'X
UNIVERSITY PARIS OUEST
F-92001 NANTERRE CEDEX
FRANCE
E-MAIL: kloppolga@math.cnrs.fr

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CENTRAL FLORIDA
ORLANDO, FLORIDA 32816-1364
USA
E-MAIL: marianna.pensky@ucf.edu