Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2013-30

Sparse High-dimensional Varying Coefficient Model : Non-asymptotic Minimax Study

O. KLOPP[1]
M. PENSKY[2]

December, 2013

1 CREST and MODAL'X, University Paris Ouest, Nanterre, France.
2 University of Central Florida, Orlando, FL 32816, USA.

Sparse high-dimensional varying coefficient model: non-asymptotic minimax study

Olga Klopp, CREST and MODAL'X, University Paris Ouest, 92001 Nanterre, France
Marianna Pensky, University of Central Florida, Orlando, FL 32816, USA

Abstract

The objective of the present paper is to develop a minimax theory for the varying coefficient model in a non-asymptotic 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.

Keywords: varying coefficient model, sparse model, minimax optimality

AMS 2000 subject classification: 62H12, 62J05, 62C20

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 [9] and Hastie and Tibshirani [14] many methods for estimation and inference in the varying coefficient model have been developed (see, e.g., [38, 15, 12, 19] for the kernel-local polynomial smoothing approach, [16, 18, 17] for the polynomial spline approach, [14, 15, 8] for the smoothing spline approach and [13] 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.* [37] proposed a new procedure based on a local rank estimator; Kai *et al.* [20] 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.* [41] studied the multivariate version of the model.

One important aspect that has not been well studied in the existing literature is the non-asymptotic approach to estimation, prediction and variables selection in the varying coefficient model. Although, in practical applications, only a finite number of measurements are available, existing methods typically provide asymptotic evaluation of the precision of estimation procedures under the assumption that the number of observations tends to infinity. Recently, few

authors used a non-asymptotic approach to the problem. Li *et al.* [23] applied group Lasso for variable selection while Lian [24] used extended Bayesian information criterion. Klopp and Pensky [21] carried out variable selection by nuclear matrix norm penalization. Fan *et al.* [11] applied nonparametric independence screening, their results were extended by Lian and Ma [25] to include rank selection in addition to variable selection.

Some interesting questions arise in this non-asymptotic 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 *non-asymptotic 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, e.g., [2], [4], [5], [35]). A related LASSO-type procedure is the group-LASSO, where the covariates are assumed to be clustered in groups (see, for example, [40, 1, 7, 27, 28, 26], and references therein).

In the present paper, we also consider the case when the solution is sparse, in particular, only 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 [24]. One important difference, however, is that in [24], the estimator is not adaptive to the smoothness of the time dependent covariates. In addition, Liang [24] 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 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 notations and assumptions on the model and provides a discussion of the assumptions. Section 3 describes the block thresholding LASSO estimator, evaluates the non-asymptotic 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 $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \ldots, n$ be sampled independently from the varying coefficient model

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

Here the noise variables ξ_i are independent and σ is known, $\mathbf{W} \in \mathbb{R}^p$ are random vectors of predictors, $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_p(\cdot))^T$ is an unknown vector-valued function of regression coefficients and $t \in [0, 1]$ is a random variable with the unknown density function g. The triple $(\mathbf{W}, t, Y)$ is independent. The goal is to estimate vector function $f(\cdot)$ on the basis of observations $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \ldots, n$.

In order to estimate $\mathbf{f}$, following Klopp and Pensky (2013), we expand it over a basis $(\phi_l(\cdot)), l = 0, 1, \ldots, \infty$, 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

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

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

$$Y_i = \mathrm{Tr}(\mathbf{A}^T\boldsymbol{\phi}(t_i)\mathbf{W}_i^T) + \mathbf{W}_i^T\boldsymbol{\rho}(t_i) + \sigma\xi_i, \quad i = 1, \ldots, n. \tag{1.3}$$

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

1.2 Tensor approach to estimation

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

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

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

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

In what follows, we denote

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

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 $i = 1, \cdots, n$, 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

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

which implies

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

Let

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

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

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

2 Assumptions and notations

2.1 Notations

In what follows, we use bold script for matrices and vectors, e.g., $\mathbf{A}$ or $\mathbf{a}$, and superscripts to denote elements of those matrices and vectors, e.g., $\mathbf{A}^{(i,j)}$ 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 $\mathbf{q}(t) = (q_1(t), \cdots, q_p(t))^T$, 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 $\mathbf{r} = (r_1, \cdots, r_p)^T$, denote $r_j^* = r_j \wedge r_j'$ and $r_{\min}^* = \min_j r_j^*$.

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 non-constant 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.

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

$$\sum_{k=0}^{l} \phi_k^2(t) \leq C_\phi^2 (l+1). \tag{2.1}$$

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

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

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

(A3). Vectors $\mathbf{W}_i$ are i.i.d copies of a random vector $\mathbf{W}$ having distribution Π on a given set of vectors $\mathcal{X}$. Here Π is such that matrix $\mathbb{E}(\mathbf{W}\mathbf{W}^T) = \boldsymbol{\Omega}$ has eigenvalues bounded above and below

$$0 < \omega_{\min} = \lambda_{\min}(\boldsymbol{\Omega}) \leq \lambda_{\max}(\boldsymbol{\Omega}) = \omega_{\max} < \infty,$$

and, for any $j = 1, \cdots, p$, $\mathbb{E}(\mathbf{W}^{(j)})^4 \leq V$. Also, for any $\mu > 0$, there exist positive constants U_μ and C_μ and a set $\mathcal{W}_\mu$ such that

$$\mathbf{W} \in \mathcal{W}_\mu \Longrightarrow (\|\mathbf{W}\|_2 \leq U_\mu) \cap \left(\max_{j=1,\cdots,p} |\mathbf{W}^{(j)}| \leq C_\mu \right), \quad \mathbb{P}\left(\mathbf{W} \in \mathcal{W}_\mu\right) \geq 1 - 2n^{-\mu}. \tag{2.2}$$

Here, $\omega_{\min} = \omega_{\min}(p)$, $\omega_{\max} = \omega_{\max}(p)$, $U_\mu = U_\mu(p)$, $C_\mu = C_\mu(p)$ and $V = V(p)$.

(A4). Functions $f_j(t)$ have efficient representations in basis ϕ_l, in particular, for any $j = 1, \cdots, p$, one has

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

for some $C_a > 0$, $1 \leq \nu_j < \infty$ 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, $j = 1, \cdots, p$, by $\boldsymbol{\nu}$ and $\mathbf{r}$, respectively, and the set of indices of finite elements r_j by $\mathcal{J}$:

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

(A5). Variables ξ_i, $i = 1, \cdots, n$, are i.i.d. *sub-Gaussian* such that

$$\mathbb{E}\xi_i = 0, \quad \mathbb{E}\xi_i^2 = 1 \quad \text{and} \quad \|\xi_i\|_{\psi_2} \leq K$$

where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm.

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 for representation of $\mathbb{E}Y$.

- Assumptions **(A1)** and **(A2)** deal 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. 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 $h = 0, 1, \cdots$, $i = 0, \cdots, 2^h - 1$. Set $\phi_0(t) = 1$ and $\phi_j(t) = \psi_{h,i}(t)$ with $j = 2^h + i + 1$. If $l = 2^J$, then $C_\phi = \|\psi\|_\infty$. Since for $2^J < l < 2^{J+1}$ one has $(l+1) \geq (2^{J+1}+1)/2$, for any l it is sufficient to choose $C_\phi = 2\|\psi\|_\infty$.

  Assumption that $\phi_{\min}$ and $\phi_{\max}$ are absolute constants independent of L is guaranteed by the fact that sampling density g is bounded above and below. For example, if $g(t) = 1$, one has $\phi_{\min} = \phi_{\max} = 1$.

- Note that since all assumptions on vectors $\mathbf{W}_i$ are satisfied up to a multiplicative constant, we assume that the expected length of the vector $\mathbf{W}$ is equal to one, i.e. $\mathbb{E}(\mathbf{W}^T\mathbf{W}) = 1$.

  Assumption **(A3)** is satisfied for many collections of vectors. In particular, it holds when $\mathbf{W}_i$ are normally distributed with zero mean vector and covariance matrix $p^{-1}\mathbf{I}$; when $\mathbf{W}_i$ are uniformly distributed on the set of canonical basis vectors $\mathbf{e}_j, j = 1, \cdots, p$; or when $\mathbf{W}_i^{(j)}$, $i = 1, \cdots, n$, $j = 1, \cdots, p$, are independent symmetric Bernoulli variables.

- Assumption **(A4)** describes sparsity of the vectors of coefficients of functions $f_j(t)$ in basis ϕ_l, $j = 1, \cdots, p$. If $\nu_j = 2$, then f_j belongs to a Sobolev ball of smoothness r_j and radius C_a. If $\nu_j < 2$, coefficients a_{jl} of f_j are sparse. For example, 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. Note that Assumption **(A4)** allows each non-constant function f_j to have its own sparsity pattern.

- Assumption **(A5)** 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. Note that $\mathbb{E}\xi_i^2 = 1$ implies that $K \leq 1$.

3 Estimation strategy and non-asymptotic 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, $j = 1, \cdots, p$, 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 $\mathbf{a}_{jl} = (a_{j,d(l-1)+1}, \cdots, a_{j,dl})^T$ the sub-vector of coefficients of function f_j in block l, $l = 1, \cdots, M$. Let K_l be the subset of indices associated with $\mathbf{a}_{jl}$. We impose block norm on matrix $\mathbf{A}$ in (1.3) 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} for the constant portions of functions f_j and l_2 norms for each of the block vectors of coefficients $\mathbf{a}_{jl}$, $j = 1, \cdots, p$, $l = 1, \cdots, M$. 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 $\mathbf{A}$ as a solution of the following convex optimization problem

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

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 $\mathbf{a} = \text{Vec}(\mathbf{A})$ as

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

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

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

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 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 $\widehat{\mathbf{A}}$ under assumptions **(A0)**–**(A5)**.

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 what follows, 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 non-constant with coefficients satisfying condition (2.3) in **(A4)**, 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 $\mathbf{f} \in \mathcal{F}_{s_0,s,\boldsymbol{\nu},\mathbf{r}}(C_a)$. Denote $r_{max} = \max\{r_j : j \in \mathcal{J}\}$ and

$$n_{low} = \frac{2\,\sigma^2\kappa}{C_a^2\,\omega_{\max}\,\phi_{\max}}\ \max\left\{1, \left(\frac{6}{s}\right)^{2r_{max}+1}\right\}, \tag{3.5}$$

$$\Delta_{lower}(s_0, s, n, \mathbf{r}) = \max\left\{\frac{\kappa\,\sigma^2\,s_0}{4n\,\omega_{\max}\,\phi_{\max}},\ \frac{1}{8}\sum_{j\in\mathcal{J}} C_a^{\frac{2}{2r_j+1}}\ \left(\frac{\sigma^2\,\kappa}{n\,\omega_{\max}\,\phi_{\max}}\right)^{\frac{2r_j}{2r_j+1}}\right\}. \tag{3.6}$$

Then, the following statement holds.

Theorem 1 *Let $s \geq 1$ and $s_0 \geq 3$. Denote by $\mathcal{F} = \mathcal{F}_{s_0,s,\boldsymbol{\nu},\mathbf{r}}(C_a)$ a class of vector functions $\mathbf{f}(t)$ satisfying conditions* **(A0)** *and* **(A4)** *in the basis ϕ_l, $l = 0, 1, \cdots$, that obeys assumptions* **(A1)** *and* **(A2)**. *Consider observations Y_i in model (1.3) with $\mathbf{W}_i$, $i = 1, \cdots, n$, such that assumption* **(A3)** *holds. Assume that, conditionally on $\mathbf{W}_i$ and t_i, the variables ξ_i are Gaussian $\mathcal{N}(0,1)$ and that $n \geq n_{low}$. Then, for any $\kappa < 1/8$ and any estimator $\tilde{\mathbf{f}}$ of $\mathbf{f}$, one has*

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

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 $n \geq n_{low}$ 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

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

3.3 Adaptive estimation and upper bounds for the risk

Denote

$$\mathbf{t} = (t_1, \cdots, t_n), \quad \mathbb{W} = (\mathbf{W}_1, \cdots, \mathbf{W}_n), \tag{3.9}$$

i.e., $\mathbb{W}$ is the $p \times n$ matrix with columns $\mathbf{W}_i$, $i = 1, \cdots, n$. Let also $\beta > 0$ be such that

$$2p(L+1) = n^\beta. \tag{3.10}$$

Note that, since we are using non-asymptotic approach, this relation just means that $\beta = \log(p(L+1))/\log n$. If, however, n is large, relation (3.10) implies that $p(L+1)$ cannot grow faster than a power of n.

In order to establish an upper bound for the minimax quadratic risk over the class $\mathcal{F} = \mathcal{F}_{s_0,s,\boldsymbol{\nu},\mathbf{r}}(C_a)$ of vector functions $\mathbf{f}(t)$, we need two kinds of large deviation results. First, we need to show that, with the high probability, the ratio between the eigenvalues of matrices $\widehat{\boldsymbol{\Sigma}}$ defined in (1.7) and $\boldsymbol{\Sigma} = \mathbb{E}\widehat{\boldsymbol{\Sigma}}$ is bounded above and below. Second, for any vector $\mathbf{a} \in \mathbb{R}^{p(L+1)}$, we need to obtain an upper bound for $|\langle \mathbf{a}, \mathbf{B}^T\boldsymbol{\xi}\rangle|$. This is accomplished by the following lemmas.

Lemma 1 *Let μ in (2.2) be large enough, so that*

$$n^{\mu-3} \geq \frac{Vp^2}{U_\mu^4 \log^2(n)} \tag{3.11}$$

and $n/\log n \geq N_0$ where, for β defined in (3.10) and for some $\gamma > 0$ and $0 < h < 1$,

$$N_0 = \frac{8C_\phi^2 U_\mu^2 (L+1)\phi_{\max}\omega_{\max}\left(2(\gamma+\beta)+5\,h/4\right)}{h^2\phi_{\min}^2\omega_{\min}^2}. \tag{3.12}$$

Let $\mathcal{W}_\mu$ be the set of points in $\mathbb{R}^p$ such that condition (2.2) holds and $\mathcal{W}_\mu^{\otimes n}$ be the direct product of n sets $\mathcal{W}_\mu$. Then,

$$\mathbb{P}\left(\left\{\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| < h\phi_{\min}\omega_{\min}\right\} \cap \mathcal{W}_\mu^{\otimes n}\right) \geq 1 - n^{-\gamma} - 2n^{-\mu+1}, \tag{3.13}$$

so that, on the set $\mathcal{W}_\mu^{\otimes n}$, with probability at least $1 - n^{-\gamma} - 2n^{-\mu+1}$, one has simultaneously

$$\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}) \geq (1-h)\phi_{\min}\omega_{\min}, \quad \lambda_{\max}(\widehat{\boldsymbol{\Sigma}}) \leq (1+h)\phi_{\max}\omega_{\max}. \tag{3.14}$$

Lemma 2 *Let vector $\boldsymbol{\alpha} \in \mathbb{R}^{p(L+1)}$ be partitioned into subgroups the way it was done for vector $\mathbf{a} = \mathrm{Vec}(\mathbf{A})$. Then, for any given $\mathbf{t}$ and $\mathbb{W}$ and any $\gamma_1 \geq 1$ one has*

$$\mathbb{P}\left(\frac{2\sigma\,\left|\langle\boldsymbol{\alpha}, \mathbf{B}^T\boldsymbol{\xi}\rangle\right|}{n} \leq \hat{\delta}\|\boldsymbol{\alpha}\|_{block}\middle|\mathbf{t}, \mathbb{W}\right) \geq 1 - \frac{p(L+1)}{n^{\gamma_1}} \tag{3.15}$$

where

$$\hat{\delta} = 2\sigma\left(1 + CK\sqrt{\gamma_1}\right)\left(n^{-1}\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})\log n\right)^{1/2}. \tag{3.16}$$

Lemma 1 ensures that the lowest eigenvalue of the regression matrix $\widehat{\boldsymbol{\Sigma}}$ is within a constant factor of the respective eigenvalue of matrix $\boldsymbol{\Sigma}$. 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 condition (3.12) on the relationship between parameters L, p and n. If matrix $\widehat{\boldsymbol{\Sigma}}$ is such that $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})$ is automatically bounded by below, assumption (3.12) can be removed.

Lemma 2 establishes an upper bound on the random term and, by applying a combination of LASSO and group LASSO arguments, allows to obtain the following upper bound for the quadratic risk. Define

$$N = \max\left\{\frac{8C_\phi^2 U_\mu^2 (L+1)\phi_{\max}\omega_{\max}\left(2(\gamma+\beta)+\frac{5\,h}{4}\right)}{h^2\phi_{\min}^2\omega_{\min}^2},\ \frac{(\gamma+\beta)U_\mu^2}{2\,g_2\omega_{\max}},\ \frac{(\gamma+\beta)}{g_2}\right\}. \tag{3.17}$$

Theorem 2 *Let* $\min_k(r_k \wedge r'_k) \geq 2$, $L+1 \geq n^{1/4}$ *and* $n/\log n \geq N$. *Suppose that conditions (3.10) and (3.11) hold. If* $\hat{\mathbf{a}}$ *is an estimator of* $\mathbf{a}$ *obtained as a solution of optimization problem (3.3) with* $\delta = \hat{\delta}$, *and vector function is recovered using (3.4), then, for any* $\gamma > 0$ *and* $\mu > \gamma + 1$, *one has*

$$\mathbb{P}\left(\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2 \leq \Delta(s_0, s, n, \mathbf{r})\right) \geq 1 - 8\, n^{-\gamma} \tag{3.18}$$

where

$$\begin{aligned} \Delta(s_0, s, n, \mathbf{r}) &= \frac{(1+h)\omega_{\max}\phi_{\max}}{(1-h)\omega_{\min}\phi_{\min}} \left\{ \frac{32\left(1 + CK^2(\gamma+\beta)\right)}{(1-h)\phi_{\min}} \, \frac{\sigma^2(s+s_0)\log n}{n\omega_{\min}} \right. \\ &+ \sum_{j\in\mathcal{J}} C_a^{2/(2r_j+1)} \left(\frac{6\sigma^2\left(1 + CK^2(\gamma+\beta)\right)}{n(1-h)\omega_{\min}\phi_{\min}} \right)^{\frac{2r_j}{2r_j+1}} (\log n)^{\frac{(2-\nu_j)_+}{\nu_j(2r_j+1)}} \\ &+ \left. \frac{C_a^2 s}{(1+h)\, n\, \phi_{\max}} \left(g_2 + (1-h)\phi_{\min}\right) \right\}. \end{aligned} \tag{3.19}$$

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. Moreover, conclusions of Theorem 2 are derived without any asymptotic assumptions on n, p and L.

In order to assess optimality of estimator $\hat{\mathbf{a}}$, we 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 (3.18) which are not bounded above and below by an absolute constant are σ, $\omega_{\min}$, $\omega_{\max}$, s and s_0. Hence, $\Delta(s_0, s, n, \mathbf{r}) \leq C\, \Delta_{upper}(s_0, s, n, \mathbf{r})$ with

$$\Delta_{upper}(s_0, s, n, \mathbf{r}) = \frac{\omega_{\max}}{\omega_{\min}} \left[\frac{\sigma^2 s_0 \log n}{n\omega_{\min}} + \sum_{j\in\mathcal{J}} \left(\frac{\sigma^2}{n\omega_{\min}} \right)^{\frac{2r_j}{2r_j+1}} (\log n)^{\frac{(2-\nu_j)_+}{\nu_j(2r_j+1)}} + \frac{s}{n} \right], \tag{3.20}$$

where C is an absolute constant independent of n, p and σ^2.

Inequality (3.20) implies that, for any values of parameters, the ratio between the upper and the lower bound for the risk (3.6) is bounded by, at most, $C \log n \omega_{\max}^2/\omega_{\min}^2$. Note that $\omega_{\max}/\omega_{\min}$ is the condition number of matrix $\boldsymbol{\Omega}$. Hence, if matrix $\boldsymbol{\Omega}$ is well conditioned, so that $\omega_{\max}/\omega_{\min}$ is bounded above by a constant, estimator $\hat{\mathbf{f}}$ attains optimal convergence rates up to, at most, a $\log n$ factor. Moreover, if s/s_0 is bounded or $n\omega_{\min}/\sigma^2$ is relatively large, then estimator $\hat{\mathbf{f}}$ attains optimal convergence rates up to a constant factor if all functions $f_j(t)$ are spatially homogeneous, i.e., if $\min_j \nu_j \geq 2$. In particular, if all functions in assumption **(A4)** belong to the same space, then the following corollary is valid.

Corollary 1 *Let conditions of Theorem 2 hold with* $r_j = r$ *and* $\nu_j = \nu$, $j - 1, \cdots, p$, *and matrix* $\boldsymbol{\Omega}$ *be well conditioned, i.e.* $\omega_{\max}/\omega_{\min}$ *is bounded by for some absolute constant independent of* n, p *and* σ. *Then,*

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

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:

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

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

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left\{ n^{-1}\|\mathbf{Y}-\mathbf{B}\mathbf{a}\|_2^2 + \delta\|\mathbf{a}\|_{block} \quad \text{s.t.} \quad \|\mathbf{a}\|^2 \le p\,C_f^2 \right\}, \tag{3.23}$$

and estimators $\hat{f}_j$ of f_j, $j=1,\cdots,p$, are constructed using formula (3.4). Choose γ in Theorem 2 large enough, so that

$$16pC_f^2 \le n^{\gamma-1}. \tag{3.24}$$

Then, the following statement is valid.

Theorem 3 *Let $r'\ge 2$, $L+1\ge n^{1/4}$ and (3.10), (3.11) and (3.17) hold. If $\hat{\mathbf{a}}$ is an estimator of $\mathbf{a}$ obtained as a solution of optimization problem (3.23) with $\delta=\hat{\delta}$ and vector function is recovered using (3.4), then, for any γ satisfying condition (3.24) and $\mu>\gamma+1$, one has*

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

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

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 $\sqrt{p}\,\mathbf{W}_i$, $i=1,\cdots,n$, be i.i.d. standard Gaussian vectors $N(\mathbf{0},\mathbf{I}_p)$. Then, $\mathbf{\Omega}=\mathbf{I}_p/p$, so that $\omega_{\min}=\omega_{\max}=1/p$. Moreover, $\sqrt{p}\,\mathbf{W}^{(j)}$ are independent standard Gaussian variables and $p\mathbf{W}_i^T\mathbf{W}_i$ are independent chi-squared variables with p degrees of freedom. Using inequality (see, e.g., [3], page 67)

$$\mathbb{P}\left(\chi_p^2 \le p+2\sqrt{px}+2x\right) \ge 1-e^{-x}, \quad x>0,$$

for any non-negative μ_1 and μ_2 and any $j=1,\cdots,p$, obtain

$$\mathbb{P}\left(\mathbf{W}_1^T\mathbf{W}_1 \le (1+\sqrt{2}\mu_1)^2\right) \ge 1-\exp(-p\mu_1^2/2), \quad \mathbb{P}(|\mathbf{W}_1^{(j)}|\le\mu_2) \ge 1-\exp(-\mu_2^2/2)/\sqrt{2\pi\mu_2}.$$

Choose $\mu>0$ and set $\mu_1^2=2\mu\log n/p$ and $\mu_2^2=2\mu\log n+2\log p$. Then, $\exp(-p\mu_1^2/2)=p\exp(-\mu_2^2/2)=n^{-\mu}$ and Assumption **(A3)** holds with

$$C_\mu=\sqrt{2\mu\log n+2\log p}, \quad U_\mu=1+\sqrt{2\mu\log n/p}.$$

Example 2 Orthonormal dictionary Let $\sqrt{p}\,\mathbf{W}_i$, $i = 1, \cdots, n$, be uniformly distributed on a set of canonical vectors $\mathbf{e}_k$, $k = 1, \cdots, p$. Then, again, $\mathbf{\Omega} = \mathbf{I}_p/p$, so that $\omega_{\min} = \omega_{\max} = 1/p$. Moreover, $\|\mathbf{W}_1\|_2^2 = 1$ and $\mathbf{W}^{(j)}$ are independent Bernoulli variables, so that $|\mathbf{W}^{(j)}| \leq 1$. Therefore, for any μ,

$$C_\mu = 1, \quad U_\mu^2 = 1,$$

and one can set $\mu = \infty$.

Example 3 Symmetric Bernoulli dictionary Let $\sqrt{p}\mathbf{W}_i^{(j)}$, $i = 1, \cdots, n$, $j = 1, \cdots, p$, be independent symmetric Bernoulli variables

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

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

$$C_\mu = 1/\sqrt{p}, \quad U_\mu^2 = 1,$$

and, same as in Example 2, one can take $\mu = \infty$.

Note that in all three cases above, N_0 in (3.12) is of the form $N_0 = Cp(L+1)$ where constant C depends on the type of matrix $\mathbb{W}$. Under conditions of Theorem 2, the upper bounds for the risk are of the form

$$\Delta_{upper}(s_0, s, n, \mathbf{r}) = C\left[\frac{\sigma^2 p s_0 \log n}{n\omega_{\min}} + \sum_{j\in\mathcal{J}}\left(\frac{p\sigma^2}{n}\right)^{\frac{2r_j}{2r_j+1}}(\log n)^{\frac{(2-\nu_j)_+}{\nu_j(2r_j+1)}} + \frac{s}{n}\right].$$

4.2 Discussion

In the present paper, we provided a non-asymptotic 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 used in the context of additive or compound functional models (see, e.g., [10] or [31]).

The adaptive estimator is obtained using block LASSO approach which can be viewed as a version of 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 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 [41]. However, this is a matter of a future investigation.

Acknowledgments

Marianna Pensky was partially supported by National Science Foundation (NSF), grant DMS-1106564. The authors want to thank Alexandre Tsybakov for extremely valuable suggestions and discussions.

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 $\mathbf{f}_{\boldsymbol{\omega}}(t) = (f_{1,\boldsymbol{\omega}}, \cdots, f_{p,\boldsymbol{\omega}})^T$ 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 $\omega_{kl} \in \{0,1\}$ for $l = l_{0k}, l_{0k}+1, \cdots, 2l_{0k}-1$, $k = 1, \cdots, p$. Let K_0 and K_1, respectively, be the sets of indices such that $K_0 \cap K_1 = \emptyset$ and $u_k = u$ if $k \in K_1$ and $u_k = 0$ otherwise, $v_k = v$ if $k \in K_0$ and $v_k = 0$ otherwise, so that $u_k v_k = 0$.

In order assumption (2.3) holds, one needs $u \leq C_a$ and

$$v^{\nu_j} \sum_{l=l_{0k}}^{2l_{0k}-1} (l+1)^{\nu_j r_j{}'} \leq (C_a)^{\nu_j}, \ j \in \Upsilon. \tag{5.2}$$

By simple calculations, it is easy to verify that condition (5.2) is satisfied if we set

$$u \leq C_a, \quad v = C_a (2l_{0k})^{-(r_k+1/2)}, \tag{5.3}$$

where the constancy of v implies that l_{0k} in (5.3) 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

$$D^2 = u^2 \sum_{k \in K_1} |\omega_{k0} - \tilde{\omega}_{k0}| + v^2 \sum_{k \in K_0} \sum_{l=l_{0k}}^{2l_{0k}-1} |\omega_{kl} - \tilde{\omega}_{kl}|. \tag{5.4}$$

Let $P_{\mathbf{f}}$ and $P_{\tilde{\mathbf{f}}}$ be probability measures corresponding to test functions $\mathbf{f}$ and $\tilde{\mathbf{f}}$, respectively. Using that, conditionally on W_i and t_i, the variables ξ_i are Gaussian $\mathcal{N}(0,1)$, 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

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

where

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

and, due to conditions **(A2)** and **(A3)**,

$$\begin{aligned} \mathbb{E} \left[Q_1(\mathbf{f}) - Q_1(\tilde{\mathbf{f}}) \right]^2 &= \mathbb{E} \left((\mathbf{f} - \tilde{\mathbf{f}})^T(t_1) W_1 W_1^T (\mathbf{f} - \tilde{\mathbf{f}})(t_1) \right) = \mathbb{E} \left((\mathbf{f} - \tilde{\mathbf{f}})^T(t_1) \, \Omega \, (\mathbf{f} - \tilde{\mathbf{f}})(t_1) \right) \\ &\leq \omega_{\max} \, \mathbb{E} \left(\left\| (\mathbf{f} - \tilde{\mathbf{f}})(t_1) \right\|_2^2 \right) \leq \omega_{\max} \, \phi_{\max} D^2, \end{aligned}$$

where D^2 is defined in (5.4).

In order to derive the lower bounds for the risk, we use Theorem 2.5 of Tsybakov (2009) which implies that, if a set Θ of cardinality $M+1$ contains sequences $\boldsymbol{\omega}_0, \cdots, \boldsymbol{\omega}_M$ with $M \geq 2$ such that, for any $j = 1, \cdots, M$, one has $\|f_{\boldsymbol{\omega}_0} - f_{\boldsymbol{\omega}_j}| \geq D > 0$, $P_{\boldsymbol{\omega}_j} << P_{\boldsymbol{\omega}_0}$ and $\mathcal{K}(P_{\mathbf{f}_j}, P_{\mathbf{f}_0}) \leq \kappa \log M$ with $0 < \kappa < 1/8$, then

$$\inf_{\tilde{\boldsymbol{\omega}}} \sup_{f_{\boldsymbol{\omega}}, \boldsymbol{\omega} \in \Theta} \mathbb{P}\left(\|\mathbf{f}_{\boldsymbol{\omega}} - \mathbf{f}_{\tilde{\boldsymbol{\omega}}}\|_2 \geq D/2\right) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\kappa - \sqrt{\frac{2\kappa}{\log M}}\right). \tag{5.5}$$

Now, we consider two separate cases.

Case 1. Let the first s_0 functions be constant and the rest of the functions be equal to identical zero. Then, $v = 0$ and $K_1 = \{1, \cdots, s_0\}$. Use the Varshamov-Gilbert Lemma (Lemma 2.9 of [34]) to choose a set Θ of ω with $\text{card}(\Theta) \geq 2^{s_0/8}$ and $D^2 \geq u^2 s_0/8$. Inequality

$$\mathcal{K}(P_{\mathbf{f}_j}, P_{\mathbf{f}_0}) \leq (2\sigma^2)^{-1} n \omega_{\max} \phi_{\max} u^2 s_0/8 \leq \kappa \log\left(\text{card}(\Theta)\right)$$

is satisfied if $u^2 = 2\,\sigma^2\kappa/(n\omega_{\max}\phi_{\max})$. Then, $D^2 = (s_0\sigma^2\kappa)/(4n\,\omega_{\max}\,\phi_{\max})$ and $u \leq C_a$ provided $n \geq 2\,\sigma^2\kappa/(C_a^2\omega_{\max}\phi_{\max})$.

Case 2. Let the first s functions be time-dependent and the rest of the functions be equal to identical zero. Then $u = 0$, v is given by formula (5.3), $K_0 = \{1, \cdots, s\}$. Let $r_k, k \in K_0$, coincide with the values of finite components of vector $\mathbf{r}$. Denote

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

Use Varshamov-Gilbert Lemma to choose a set Θ of ω with $\text{card}(\Theta) \geq 2^{\mathcal{L}/8}$ 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

$$v^2 \leq (\sigma^2\kappa)/(4n\,\omega_{\max}\,\phi_{\max}),$$

which, together with (5.3) and (5.4) imply that

$$l_{0k} = \left\lfloor \frac{1}{2}\left(\frac{4\,C_a^2\,n\omega_{\max}\phi_{\max}}{\sigma^2\kappa}\right)^{\frac{1}{2r_k+1}} \right\rfloor + 1, \quad D^2 \geq \frac{C_a^2}{16}\sum_{k=1}^{s}\left(\frac{4\,C_a^2\,n\omega_{\max}\phi_{\max}}{\sigma^2\kappa}\right)^{-\frac{2r_k}{2r_k+1}}$$

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

5.2 Proofs of the large deviation inequalities

Proof of Lemma 1. Let $\mathcal{W}_\mu$ be the set of points described in condition (2.2) of Assumption **(A3)**. Denote the direct product of n sets $\mathcal{W}_\mu$ by $\mathcal{W}_\mu^{\otimes n}$. Then,

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

Consider random matrices

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

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 [33]:

Proposition 1 (Theorem 1.6, Tropp (2011)) *Let $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$ be independent random matrices in $\mathbb{R}^{m_1 \times m_2}$ such that $\mathbb{E}(\boldsymbol{\zeta}_i) = 0$. Define*

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

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

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\zeta}_i \right\| \leq 2 \max\left\{ \sigma_{\boldsymbol{\zeta}} \sqrt{\frac{t + \log(d)}{n}}, T\, \frac{t + \log(d)}{n} \right\}, \tag{5.6}$$

where $d = m_1 + m_2$.

In order to find $\sigma_{\boldsymbol{\zeta}}$, note that

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

and, similarly,

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

Here,

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

and

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

so that

$$\sigma_{\boldsymbol{\zeta}}^2 \leq 2\, C_\phi^2\, U_\mu^2\, (L+1)\phi_{\max}\omega_{\max}. \tag{5.7}$$

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

$$T = \sup \|\boldsymbol{\zeta}_1\| \leq 2 \sup \|\boldsymbol{\Sigma}_1 \mathbb{I}(\mathcal{W}_\mu))\| = 2 \sup \|\boldsymbol{\Omega}_1 \mathbb{I}(\mathcal{W}_\mu)\| \|\boldsymbol{\Phi}_1\| \leq 2\, C_\phi^2 U_\mu^2 (L+1). \tag{5.8}$$

Apply Bernstein inequality (5.6) with $\sigma_{\boldsymbol{\zeta}}^2$ and T given by formulae (5.7) and (5.8), respectively. Then, using (3.10) and $t = \gamma \log n$, obtain for any $\gamma > 0$, with probability at least $1 - n^{-\gamma}$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\zeta}_i \right\| \leq 4 \max\left\{ \frac{C_\phi U_\mu \sqrt{(L+1)\phi_{\max}\omega_{\max}(\gamma+\beta)\log n}}{\sqrt{n}}, \frac{C_\phi^2 U_\mu^2 (L+1)(\gamma+\beta)\log n}{n} \right\}. \tag{5.9}$$

Now, in order to apply inequality (5.9) to $\mathbf{Z}_i$, observe that $\mathbf{Z}_i - \boldsymbol{\zeta}_i = \boldsymbol{\Sigma}_i \mathbb{I}(\mathcal{W}_\mu^c) - \mathbb{E}(\boldsymbol{\Sigma}_i \mathbb{I}(\mathcal{W}_\mu^c)$ and

$$\begin{aligned} \|\mathbb{E}(\boldsymbol{\Sigma}_i \mathbb{I}(\mathcal{W}_\mu^c))\|^2 &= \|\mathbb{E}\left(\boldsymbol{\Omega}_i \mathbb{I}(\mathcal{W}_\mu^c)\right) \otimes \mathbb{E}(\boldsymbol{\Phi}_i)\|^2 \leq \|\mathbb{E}\left(\boldsymbol{\Omega}_i \mathbb{I}(\mathcal{W}_\mu^c)\right)\|^2 \|\mathbb{E}(\boldsymbol{\Phi}_i)\|^2 \\ &\leq \left(\mathbb{E}\|\boldsymbol{\Omega}_i \mathbb{I}(\mathcal{W}_\mu^c)\|\right)^2 \mathbb{E}\|\boldsymbol{\Phi}_i\|_2^2 \leq V p^2 \, n^{1-\mu} C_\phi^4 (L+1)^2. \end{aligned} \tag{5.10}$$

due to the fact that $\mathbb{E}(\mathbf{W}^{(j)})^4 \leq V$ for any $j = 1, \cdots, p$. Hence, combining (5.9) and (5.10), we derive

$$\mathbb{P}\left(\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\right\| < z\right\} \cap \left\{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\right\}\right) \geq \mathbb{P}\left(\left\{\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\zeta}_i\right\| < z - \|\mathbb{E}(\boldsymbol{\Sigma}_1 \mathbb{I}(\mathcal{W}_\mu^c))\|\right\} \cap \left\{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\right\}\right)$$
$$\geq \mathbb{P}\left(\left\{\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\zeta}_i\right\| < z - C_\phi^2 (L+1)\, p \sqrt{V\, n^{1-\mu}}\right\} \cap \left\{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\right\}\right) \geq 1 - n^{-\gamma} - 2n^{-\mu+1}$$

for any z such that

$$\begin{aligned} z &\geq 4 \max\left\{\frac{C_\phi U_\mu \sqrt{(L+1)\phi_{\max}\omega_{\max}(\gamma+\beta)\log n}}{\sqrt{n}}, \ \frac{C_\phi^2 U_\mu^2 (L+1)(\gamma+\beta)\log n}{n}\right\} \\ &+ C_\phi^2 (L+1)\, p \sqrt{V\, n^{1-\mu}}. \end{aligned} \tag{5.11}$$

Note that under condition (3.11), one has

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

It is easy to check that, whenever $n \geq N_0$ where N_0 is defined in (3.12), condition (5.11) is satisfied with

$$z = 4 C_\phi U_\mu \sqrt{n^{-1}(L+1)\phi_{\max}\omega_{\max}(\gamma+\beta)\log(n)} + 5\, C_\phi^2 U_\mu^2 (L+1) n^{-1} \leq h\, \omega_{\min}\, \phi_{\min},$$

which implies that

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leq h \omega_{\min}\phi_{\min}\right) \geq 1 - 2n^{1-\mu} - n^{-\gamma}. \tag{5.12}$$

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

Proof of Lemma 2. Note that

$$\begin{aligned} n^{-1/2} \left|\langle \boldsymbol{\alpha}, \mathbf{B}^T \boldsymbol{\xi}\rangle\right| &\leq n^{-1/2} \max_{1\leq j\leq p} |(\mathbf{B}^T\boldsymbol{\xi})^{(j,0)}| \sum_{j=1}^p |\alpha_{j0}| \\ &+ n^{-1/2} \max_{\substack{1\leq j\leq p \\ 1\leq l\leq M}} \sqrt{\sum_{k\in K_l} \left[(\mathbf{B}^T\boldsymbol{\xi})^{(j,k)}\right]^2} \sum_{j=1}^p \sum_{l=1}^M \sqrt{\sum_{k\in K_l} \alpha_{jk}^2}. \end{aligned} \tag{5.13}$$

Fix vectors $\mathbf{W}_1, \cdots, \mathbf{W}_n$ and $\mathbf{t}$. Using Hoeffding-type inequality for sub-gaussian random variables (see, e.g., Proposition 5.2 in [36]) we obtain that, for any $\gamma_1 > 0$,

$$\mathbb{P}\left(n^{-1/2} \max_{1\leq j\leq p} |(\mathbf{B}^T\boldsymbol{\xi})^{(j,0)}| \leq \sqrt{C\, K^2 \gamma_1\, \lambda_{\max}(\widehat{\boldsymbol{\Sigma}})\, \log(n)} \ \middle| \mathbf{t}, \mathbb{W}\right) \geq 1 - e\, p\, n^{-\gamma_1} \tag{5.14}$$

where $C > 0$ is an absolute constant.

For the second maximum in (5.13), we use a corollary of the Hanson-Wright inequality for sub-gaussian random vectors (see Theorem 2.1 in [30]), which states that, for any matrix $\mathbf{Q} \in \mathbb{R}^{n\times d}$ and any $\gamma_1 > 0$, one has

$$\mathbb{P}\left(\|\mathbf{Q}\boldsymbol{\xi}\|_2 \le \|\mathbf{Q}\|_2 + C\,K^2\lambda_{\max}(\mathbf{Q})\sqrt{\gamma_2\log(n)}\right) \ge 1 - 2n^{-\gamma_1}.$$

Applying this inequality with matrix $\mathbf{Q} = n^{-1/2}\,\mathbf{B}_{j,l}$, where $\mathbf{B}_{j,l}$ is the $n \times d$ sub-matrix of matrix $\mathbf{B}$, and observing that $\lambda_{\max}(\mathbf{Q}) \le \sqrt{\lambda_{\max}(\widehat{\mathbf{\Sigma}})}$, $\|\mathbf{Q}\|_2^2 \le d\lambda_{\max}(\widehat{\mathbf{\Sigma}})$ and $d \approx \log n$, obtain for any $\gamma_1 > 0$

$$\mathbb{P}\left(n^{-1/2}\max_{\substack{1\le j\le p\\ 1\le l\le M}}\sum_{k\in K_l}\left[(\mathbf{B}^T\boldsymbol{\xi})^{(j,k)}\right] \le (1 + CK^2\sqrt{\gamma_2})\sqrt{\lambda_{\max}(\widehat{\mathbf{\Sigma}})\,\log(n)}\,\middle|\,\mathbf{t},\mathbb{W}\right) \ge 1 - \frac{2\,pL}{n^{\gamma_1}\,\log(n)} \tag{5.15}$$

where we used $M = L/\log(n)$.

5.3 Proofs of the upper bounds for the risk

Proof of Theorem 2. The proof is based on fairly standard LASSO theory arguments and two supplementary lemmas, Lemma 3 and Lemma 4, formulated and proved in Section 5.4.

Fix $\mathbf{t}$ and vectors $\mathbf{W}_j$, $j = 1, \cdots, p$. Set $\gamma_1 = \gamma + \beta$ and consider a set Ξ of values of the vector $\boldsymbol{\xi}$ such that

$$2\sigma n^{-1}\left|\langle\hat{\mathbf{a}} - \boldsymbol{\alpha}, \mathbf{B}^T\boldsymbol{\xi}\rangle\right| \le \hat{\delta}\|\hat{\mathbf{a}} - \boldsymbol{\alpha}\|_{block} \quad \text{for} \quad \boldsymbol{\xi} \in \Xi, \tag{5.16}$$

and $\mathbb{P}(\Xi) \ge 1 - \frac{p(L+1)}{n^{\gamma_1}} = 1 - n^{-\gamma}$ where $\hat{\delta}$ is given by formula (3.16). This set Ξ exists according to Lemma 2.

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

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

Using (1.4), for $\boldsymbol{\xi} \in \Xi$, one obtains

$$\frac{\|\mathbf{B}(\hat{\mathbf{a}} - \mathbf{a})\|_2^2}{n} \le \frac{3\,\|\mathbf{B}(\boldsymbol{\alpha} - \mathbf{a})\|_2^2}{n} + \frac{8\|\mathbf{b}\|_2^2}{n} - 2\delta\|\hat{\mathbf{a}}\|_{block} + 2\delta\|\boldsymbol{\alpha}\|_{block} + 2\hat{\delta}\|\hat{\mathbf{a}} - \boldsymbol{\alpha}\|_{block}. \tag{5.17}$$

Denote by $\mathcal{J}$ the set of indices corresponding to non-constant functions and $\mathcal{J}^c$ its complementary set. For $1 \le j \le p$ consider sets

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

It is easy to see that $l \in G_{j0}$ whenever $j \in \mathcal{J}^c$. Choose $\alpha_{j0} = a_{j0}$ if $a_{j0} \ne 0$, $\alpha_{j0} = 0$ otherwise, and $\boldsymbol{\alpha}_{jl} = \mathbf{a}_{jl}$ if $j \in \mathcal{J}$ and $l \in G_{j1}$ and $\boldsymbol{\alpha}_{jl} = \mathbf{0}$ otherwise. Set $\delta = \hat{\delta}$. Then, using the fact that $\lambda_{\min}(\widehat{\mathbf{\Sigma}})\|\mathbf{c}\|_2 \le n^{-1}\|\mathbf{B}\mathbf{c}\|_2^2 \le \lambda_{\max}(\widehat{\mathbf{\Sigma}})\|\mathbf{c}\|_2$ for any vector $\mathbf{c}$, rewrite inequality (5.17) as

$$\begin{aligned} \lambda_{\min}(\widehat{\mathbf{\Sigma}})\|(\hat{\mathbf{a}} - \mathbf{a})\|_2^2 \;&\le\; 3\,\lambda_{\max}(\widehat{\mathbf{\Sigma}})\|(\boldsymbol{\alpha} - \mathbf{a})\|_2^2 + 8n^{-1}\|\mathbf{b}\|_2^2 \\ &+\; 2\hat{\delta}\sum_{j\in G_{01}}|\hat{a}_{j0} - a_{j0}| + 2\hat{\delta}\sum_{j\in\mathcal{J}}\sum_{l\in G_{j1}}\|\hat{\mathbf{a}}_{jl} - \mathbf{a}_{jl}\|_2. \end{aligned} \tag{5.18}$$

Note that

$$2\hat{\delta}\sum_{j\in\mathcal{J}}\sum_{l\in G_{j1}}\|\hat{\mathbf{a}}_{jl}-\mathbf{a}_{jl}\|_2\le\frac{1}{2}\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})\sum_{j\in\mathcal{J}}\left[\sum_{l\in G_{j1}}\|\hat{\mathbf{a}}_{jl}-\mathbf{a}_{jl}\|_2^2+\frac{8\hat{\delta}^2\mathrm{card}(G_{1j})}{\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\right],$$

and similar inequality applies to the first sum in (5.18). By subtracting $0.5\,\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2$ from both sides of (5.18) and plugging in the values of $\hat{\delta}$ and $\boldsymbol{\alpha}$, derive for $\boldsymbol{\xi}\in\Xi$ and any given $(\mathbf{t},\mathbb{W})$:

$$\begin{aligned}\|(\hat{\mathbf{a}}-\mathbf{a})\|_2^2 &\le \frac{6\,\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})}{\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\left[\sum_{j=1}^{p}\sum_{l\in G_{j0}}\|\mathbf{a}_{jl}\|_2^2+\frac{8\|\mathbf{b}\|_2^2}{3\,n\,\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})}\right. \\ &+ \left.\frac{16\sigma^2\left(1+CK^2(\gamma+\beta)\right)(s_0+s)\log n}{3\,n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}+\sum_{j\in\mathcal{J}}\frac{16\sigma^2\left(1+CK^2(\gamma+\beta)\right)\mathrm{card}(G_{j1})\log n}{3\,n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\right].\end{aligned} \tag{5.19}$$

Choose sets G_{j1} so that $l\in G_{j1}$ iff $\|\mathbf{a}_{jl}\|_2^2>\varepsilon=\dfrac{6\sigma^2\left(1+CK^2(\gamma+\beta)\right)\log n}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}$ and observe that for any $1\le j\le p$ one has

$$\sum_{l\in G_{j0}}\|\mathbf{a}_{jl}\|_2^2+\frac{8\sigma^2(1+CK\sqrt{\gamma+\beta})^2\mathrm{card}(G_1j)\log n}{3\,n\,\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\le\sum_{l=1}^{M}\min\left(\|\mathbf{a}_{jl}\|_2^2,\frac{3\sigma^2(1+CK\sqrt{\gamma+\beta})^2\log n}{n\,\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\right).$$

Application of inequality (5.22) with $\varepsilon=\dfrac{6\sigma^2\left(1+CK^2(\gamma+\beta)\right)\log n}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}$ (see Lemma 3) yields that, for any given $(\mathbf{t},\mathbb{W})$ and $\boldsymbol{\xi}\in\Xi$

$$\begin{aligned}\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2 &\le \frac{\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})}{\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\left[\frac{32\sigma^2\left(1+CK^2(\gamma+\beta)\right)(s_0+s)\log n}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}+\frac{16\,\|\mathbf{b}\|_2^2}{\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})n}\right. \\ &+ \left.\sum_{j\in\mathcal{J}}C_a^{2/(2r_j+1)}\left(\frac{6\sigma^2\left(1+CK^2(\gamma+\beta)\right)}{n\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}\right)^{\frac{2r_j}{2r_j+1}}(\log n)^{\frac{(2-\nu_j)_+}{\nu_j(2r_j+1)}}\right]\end{aligned} \tag{5.20}$$

Now, consider a set $\mathcal{F}_1\subseteq\mathcal{W}_\mu^{\otimes n}$ such that (3.13) and (3.14) hold for $\mathbf{t},\mathbb{W}\in\mathcal{F}_1$ and a set $\mathcal{F}_2\subseteq\mathcal{W}_\mu^{\otimes n}$ such that, by Lemma 4, one has $n^{-1}\|\mathbf{b}\|_2^2\le g_2\,C_a^2C_\phi^2s\omega_{\max}(L+1)^{-4}$. Let $\mathcal{F}=\mathcal{F}_1\cap\mathcal{F}_2$.

Denote $r_j^*=r_j\wedge r_j'$ and $r_{\min}^*=\min_j r_j^*$. Choose $L+1\ge n^{1/4}$ and note that $r_{\min}^*\ge 2$. Using (5.24) we obtain $\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2\le\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2+C_a^2s(L+1)^{-2r_{\min}^*}$, which implies

$$\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2\le\|\hat{\mathbf{a}}-\mathbf{a}\|_2^2+C_a^2s\,n^{-1}.$$

Now, (5.20) together with Lemmas 1,2 and 4 imply

$$\mathbb{P}\left(\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2\le\Delta(s_0,s,n,\mathbf{r})\right)\ge 1-n^{-\gamma}-\mathbb{P}(\mathcal{F})\ge 1-8\,n^{-\gamma}.$$

Proof of Theorem 3. Let sets Ξ and $\mathcal{F}_1$ be such that (5.20), (3.13) and (3.14) hold and denote $\tilde{\mathcal{F}} = \Xi \cap \mathcal{F}_1$. Then, $P(\tilde{\mathcal{F}}) \geq 1 - 4n^{-\gamma}$ and (5.20) yields

$$\begin{aligned}
\mathbb{E}\|\hat{\mathbf{a}} - \mathbf{a}\|_2^2 &\leq \mathbb{E}\left[\|\hat{\mathbf{a}} - \mathbf{a}\|_2^2 \mathbb{I}(\tilde{\mathcal{F}})\right] + \mathbb{E}\left[\|\hat{\mathbf{a}} - \mathbf{a}\|_2^2 \mathbb{I}(\tilde{\mathcal{F}}^C)\right] \\
&\leq C\,\frac{\omega_{\max}}{\omega_{\min}}\left[\frac{\sigma^2 s_0 \log n}{n\omega_{\min}} + s\left(\frac{\sigma^2}{n\omega_{\min}}\right)^{\frac{2}{2r+1}} (\log n)^{\frac{(2-\nu)_+}{\nu(2r+1)}} + \frac{s\,\mathbb{E}\|\mathbf{b}\|_2^2)}{n\omega_{\max}}\right] + 16\,p\,C_f^2\,n^{-\gamma} \\
&\leq C\,\Delta_{upper}(s_0, s, n, r),
\end{aligned}$$

due to (3.24), (5.27) and $(L+1) \geq n^{1/4}$.

5.4 Proofs of supplementary lemmas

Lemma 3 *Let function $f(t) = \sum_{k=1}^{\infty} a_k \phi_k(t)$ satisfy condition* **(A4)**, *i.e.*

$$\sum_{k=0}^{\infty} |a_k|^{\nu} (k+1)^{\nu r'} \leq C_a^{\nu}, \quad r' = r + 1/2 - 1/\nu, \tag{5.21}$$

for some $C_a > 0$, $1 \leq \nu < \infty$ and $r > \min(1/2, 1/\nu)$. Let $\mathbf{a}_l$ be blocks of coefficients a_k of length d, so that $\mathbf{a}_l = (a_{(l-1)d+1}, \cdots, a_{ld})$. Then, for any $\varepsilon > 0$ one has

$$\sum_{l=0}^{\infty} \min(\|\mathbf{a}_l\|_2^2, \varepsilon d) \leq C_a^{\frac{2}{2r+1}} \varepsilon^{\frac{2r}{2r+1}} d^{\frac{(2-\nu)_+}{\nu(2r+1)}}. \tag{5.22}$$

Moreover, if $r' \geq 2$ and basis $\{\phi_k\}$ satisfies assumption **(A1)**, *then*

$$\|f - f_J\|_{\infty} \leq C_a C_{\phi} (J+1)^{-(r^* - 1/2)} \quad \textit{with} \quad f_J(t) = \sum_{k=1}^{J} a_k \phi_k(t). \tag{5.23}$$

Here, $r^* = \min(r, r')$, $(x)_+ = x$ if $x > 0$ and zero otherwise.

Proof. First, let us show that for $J \geq 1$

$$\sum_{k=J+1}^{\infty} a_k^2 \leq C_a^2 (J+1)^{-2r^*}. \tag{5.24}$$

Indeed, if $\nu \geq 2$, the Cauchy inequality yields

$$\sum_{k=J+1}^{\infty} a_k^2 \leq \left(\sum_{k=J+1}^{\infty} |a_k|^{\nu} k^{r'\nu}\right)^{2/\nu} \left(\sum_{k=J+1}^{\infty} k^{-\frac{2r'\nu}{\nu-2}}\right)^{1-2/\nu} \leq C_a^2 \left(\frac{\nu-2}{2r\nu}\right)^{\nu/2-1} (J+1)^{-2r}.$$

Since $(\nu-2)/(2r\nu) < 1$ for $\nu \geq 2$, inequality (5.24) holds. If $1 \leq \nu < 2$, then

$$\sum_{k=J+1}^{\infty} a_k^2 \leq \left(\max_{k \geq J+1} |a_k|\right)^{2-\nu} (J+1)^{-r'\nu} \left(\sum_{k=J+1}^{\infty} |a_k|^{\nu} k^{r'\nu}\right) \leq C_a^2 (J+1)^{-2r'},$$

so that (5.24) is valid.

Now, using (5.24), we prove (5.22). Again, we consider cases $\nu \geq 2$ and $1 \leq \nu < 2$, separately. If $\nu \geq 2$, then partitioning the sum into the portion for $l \leq J$ and $l > J$ (which corresponds to $k > Jd$), we derive

$$\sum_{l=1}^{\infty} \min(\|\mathbf{a}_l\|_2^2, \varepsilon d) \quad \leq \quad J\varepsilon d + C_a^2 (Jd)^{-2r}.$$

Minimizing the last expression with respect to J, we obtain (5.22) without the log-factor. If $1 \leq \nu < 2$, then

$$\sum_{l=1}^{\infty} \min(\|\mathbf{a}_l\|_2^2, \varepsilon d) \quad \leq \quad J\varepsilon d + (d\varepsilon)^{1-\nu/2} \sum_{l=J+1}^{\infty} \|\mathbf{a}_l\|_2^{\nu}.$$

Since for $1 \leq \nu < 2$

$$\|\mathbf{a}_l\|_2^2 = \sum_{k=(l-1)d+1}^{ld} a_k^2 \quad \leq \quad \left(\sum_{k=(l-1)d+1}^{ld} |a_k|^{\nu} \right)^{2/\nu},$$

one has

$$\sum_{l=J+1}^{\infty} \|\mathbf{a}_l\|_2^{\nu} \leq \sum_{k=Jd+1}^{\infty} |a_k|^{\nu} \quad \leq \quad C_a^{\nu} (Jd+1)^{-r'\nu}$$

and

$$\sum_{l=1}^{\infty} \min(\|\mathbf{a}_l\|_2^2, \varepsilon d) \quad \leq \quad J\varepsilon d + C_a^{\nu} (d\varepsilon)^{1-\nu/2} (Jd)^{-r'\nu}.$$

Minimization of the last expression with respect to J yields (5.22).

In order to prove (5.23), observe that for any $J > 1$ and $r^* > 3/2$, one has

$$\begin{aligned} \|f - f_J\|_{\infty} \quad &\leq \quad \sup_{t \in [0,1]} \sum_{l=1}^{\infty} \sqrt{\sum_{k=Jl}^{J(l+1)-1} a_k^2} \sqrt{\sum_{k=Jl}^{J(l+1)-1} \phi_k^2(t)} \leq C_a C_{\phi} \sum_{l=1}^{\infty} \sqrt{Jl} (Jl+1)^{-r^*} \\ &\leq \quad C_a C_{\phi} J^{-(r^*-1/2)} \end{aligned}$$

which completes the proof.

Lemma 4 *Let $r_j^* = r_j \wedge r_j'$ and $r_{\min}^* = \min_j r_j^* \geq 2$ in assumption* **(A4)** *and*

$$\frac{n}{\log n} \geq \max \left\{ \frac{(\gamma+\beta) U_{\mu}^2}{2\, g_2 \omega_{\max}}, \frac{(\gamma+\beta)}{g_2} \right\}. \tag{5.25}$$

Then, for any γ such that condition (3.12) holds, one has

$$\mathbb{P}\left(\left\{ n^{-1} \|\mathbf{b}\|_2^2 \leq g_2\, C_a^2 C_{\phi}^2\, s\, \omega_{\max} (L+1)^{-4} \right\} \cap \left\{ \mathbb{W} \in \mathcal{W}_{\mu}^{\otimes n} \right\} \right) \geq 1 - 2n^{-\gamma} - 2n^{-\mu+1}. \tag{5.26}$$

Here $\mathbf{b}$ is the vector with components $\mathbf{b}_i = \mathbf{W}_i^T \boldsymbol{\rho}(t_i)$, $i = 1, \cdots, n$, where $\boldsymbol{\rho}(t) = (\rho_1(t), \cdots, \rho_p(t))^T$ and $\rho_i(t)$ are defined in (1.2).

Proof. Introduce i.i.d. random variables $\beta_i = b_i^2 - \mathbb{E}b_i^2$ with $\mathbb{E}\beta_i = 0$. By direct calculations, it is easy to check that, for any i,

$$\mathbb{E}b_i^2 \;=\; \mathbb{E}\left(n^{-1}\|\mathbf{b}\|_2^2\right) = \sum_{j_1,j_2=1}^{p} \mathbf{\Omega}^{(j_1,j_2)} \int_0^1 \rho_{j_1}(t)\rho_{j_2}(t)g(t)dt \le g_2\lambda_{\max}(\mathbf{\Omega}) \int_0^1 \sum_{j=1}^{p} \rho_j^2(t)dt,$$

so Lemma 3 yields

$$\mathbb{E}\left(n^{-1}\|\mathbf{b}\|_2^2\right) \;\le\; g_2 C_a^2 C_\phi^2\, \omega_{\max} s(L+1)^{-2r^*_{\min}+1}, \tag{5.27}$$

$$\sigma_b^2 = \mathbb{E}\beta_i^2 \;\le\; \mathbb{E}b_i^4 \le g_2 C_a^4 C_\phi^4 s^2 \omega_{\max}^2 (L+1)^{-(4r^*_{\min}-2)}. \tag{5.28}$$

It is also easy to see that, for any i and $\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}$, by (5.23), one has,

$$U_b = \max(b_i^2) \le \max_t \|\rho(t)\|_2^2\ \lambda_{\max}(\mathbf{W}_i\mathbf{W}_i^T) \le C_a^2 C_\phi^2 C_{r,\nu} U_\mu^2\ s(L+1)^{-(2r^*_{\min}-1)}. \tag{5.29}$$

Now, applying Bernstein inequality to $\tilde{\beta}_i = \beta_i \mathbb{I}(\mathcal{W}_\mu) - \mathbb{E}\left(\beta_i \mathbb{I}(\mathcal{W}_\mu)\right), i = 1, \cdots, n$ with σ_b^2 and U_b given by (5.28) and (5.29), respecively, and $t = \gamma \log(n)$, one obtains, for any $\gamma > 0$, with probability at least $1 - n^\gamma$

$$\left|\frac{1}{n}\sum_{i=1}^{n} \tilde{\beta}_i\right| \le 2\max\left\{\sigma_b\sqrt{\frac{(\gamma+\beta)\log n}{n}},\, U_b\frac{(\gamma+\beta)\log n}{n}\right\}. \tag{5.30}$$

Since $\beta_i - \tilde{\beta}_i = \beta_i \mathbb{I}(\mathcal{W}_\mu^c) - \mathbb{E}\left(\beta_i \mathbb{I}(\mathcal{W}_\mu^c)\right)$, derive

$$\mathbb{P}\left(\left\{n^{-1}\|\mathbf{b}\|_2^2 < z\right\} \cap \left\{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\right\}\right) \ge \mathbb{P}\left(\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\beta}_i\right| < z - 2\mathbb{E}\left(n^{-1}\|\mathbf{b}\|_2^2\right)\right\} \cap \left\{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\right\}\right)$$

$$\ge \mathbb{P}\left(\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\zeta}_i\right| < z - 2\, g_2 C_a^2 C_\phi^2\, \omega_{\max} s(L+1)^{-2r^*_{\min}+1}\right\} \cap \left\{\mathbb{W} \in \mathcal{W}_\mu^{\otimes n}\right\}\right) \ge 1 - n^{-\gamma} - 2n^{-\mu+1}$$

for any z such that

$$z \ge 2\max\left\{\sigma_b\sqrt{\frac{(\gamma+\beta)\log n}{n}},\, U_b\frac{(\gamma+\beta)\log n}{n}\right\} + 2\, g_2 C_a^2 C_\phi^2\, \omega_{\max} s(L+1)^{-2r^*_{\min}+1}.$$

For n satisfying (5.25) one can choose

$$z = 4\, g_2 C_a^2 C_\phi^2\, \omega_{\max} s(L+1)^{-2r^*_{\min}+1}.$$

References

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

[2] Bickel, P.J., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**(4), 1705 - 1732.

[3] Birgè, L., Massart, P. (2007) Minimal Penalties for Gaussian Model Selection. *Probab. Theory Related Fields*, **138** , 33-73

[4] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Aggregation for Gaussian regression. *Ann. Statist.*, **35(4)**, 1674 - 1697.

[5] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, **1**, 169 - 194.

[6] Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer.

[7] Chesneau, C. and Hebiri, M.(2008) Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.*, **17(4)**, 317 - 326.

[8] 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.

[9] Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991) Local regression models. *Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309-376. Wadsworth and Books, Pacific Grove.

[10] Dalalyan, A., Ingster, Y., Tsybakov, A.B. (2013) Statistical inference in compound functional models. *Probab. Theory Rel. Fields*, to appear.

[11] Fan, J., Ma, Y., and Dai, W. (2013) Nonparametric Independence Screening in Sparse Ultra-High Dimensional Varying Coefficient Models. `arxiv:1303.0458v1`

[12] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.*, **27**, 1491-1518.

[13] Fan, J., and Zhang, W. (2008) Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1**, 179-195.

[14] Hastie, T.J. and Tibshirani, R.J. (1993) Varying-coefficient models. *J. Roy. Statist. Soc. B.* (Chambers, J.M. and Hastie, T.J., eds), **55** 757-796.

[15] Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Non-parametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

[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.

[17] Huang, J. Z. and Shen, H. (2004). Functional coefficient regression models for nonlinear time series: A polynomial spline approach. *Scandinavian Journal of Statistics*, **31**, 515-534.

[18] Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, **14**, 763-788.

[19] Kauermann, G. and Tutz, G. (1999). On model diagnostics using varying coefficient models. *Biometrika*, **86**, 119-128.

[20] Kai, B., Li, R., and Zou, H. (2011) New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Ann. Stat.*, **39**, 305-332.

[21] Klopp, O., Pensky, M. (2013) Non-asymptotic approach to varying coefficient model. *Electronic Journal of Statistics*, **7**, 454-479.

[22] Lee, Y.K., Mammen, E., and Park, B.U. (2012) Flexible generalized varying coefficient regression models. *Ann. Stat.*, **40**, 1906-1933.

[23] Li, G., Xue, L., and Lian,H. (2011) Semi-varying coefficient models with a diverging number of components. *Journ. of Multivar. Anal.*, **102**, 1166-1174.

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

[25] Lian, H., and Ma, S. (2013) Reduced-rank Regression in Sparse Multivariate Varying-Coefficient Models with High-dimensional Covariates. `arxiv:1309.6058v1`

[26] Lounici, K., Pontil, M., Tsybakov, A. and van de Geer, S. (2010) Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, **39(4)**, 2164-2204.

[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(1)**, 53 - 71.

[28] Meier, L., van de Geer, S. and Bühlmann, P. (2009) High-dimensional additive modeling. *Ann. Statist.*, **37(6B)**, 3779 - 3821.

[29] Mallat, S. (2009) *A Wavelet Tour of Signal Processing,* Third Ed., Elsevier, New York

[30] Rudelson, M., and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. `arXiv:1306.2872`

[31] Raskutti, G., Wainwright, M.J., Yu, B. (2012) Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journ. Machine Learning Research*, **13**, 389-427.

[32] Senturk, D. and Mueller, H. G. (2010) Functional varying coefficient models for longitudinal data. *J. Amer. Statist. Assoc.*, **105**, 1256-1264.

[33] Tropp, J.A. (2011) User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, **11(4)**.

[34] Tsybakov, A. (2009) *Introduction to Nonparametric Estimation*, Springer Series in Statistics.

[35] van de Geer, S. (2008) High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, **36**, 614 - 645.

[36] Vershynin, R. (2012) *Introduction to the non-asymptotic analysis of random matrices.* In *Compressed Sensing, Theory and Applications*, ed. Y. Eldar and G. Kutyniok, Chapter 5. Cambridge University Press.

[37] Wang, L., Kai, B., and Li, R. (2009) Local Rank Inference for Varying Coefficient Models. *J. Amer. Statist. Assoc.*, **104**, 1631-1645.

[38] 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.

[39] Yang, L., Park, B.U., Xue, L. and Hardle, W. (2006) Estimation and Testing for Varying Coefficients in Additive Models With Marginal Integration. *J. Amer. Statist. Assoc.*, **101**, 1212-1227

[40] Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68(1)** ,49 - 67.

[41] Zhu, H., Li, R., and Kong, L. (2012) Multivariate varying coefficient model for functional responses. *Ann. Stat.*, **40**, 2634-2666