1-1-2013

Non-asymptotic approach to varying coefficient model

Olga Klopp

Marianna Pensky
University of Central Florida

Electronic Journal of Statistics
Vol. 7 (2013) 454–479
ISSN: 1935-7524
DOI: 10.1214/13-EJS778

Non-asymptotic approach to varying coefficient model

Olga Klopp

MODAL'X, University Paris Ouest Nanterre, 92001 Nanterre, France
e-mail: kloppolga@math.cnrs.fr

and

Marianna Pensky[*]

Department of Mathematics, University of Central Florida,
Orlando, FL 32816-1364, USA
e-mail: marianna.pensky@ucf.edu

Abstract: In the present paper we consider the varying coefficient model which represents a useful tool for exploring dynamic patterns in many applications. Existing methods typically provide asymptotic evaluation of precision of estimation procedures under the assumption that the number of observations tends to infinity. In practical applications, however, only a finite number of measurements are available. In the present paper we focus on a non-asymptotic approach to the problem. We propose a novel estimation procedure which is based on recent developments in matrix estimation. In particular, for our estimator, we obtain upper bounds for the mean squared and the pointwise estimation errors. The obtained oracle inequalities are non-asymptotic and hold for finite sample size.

AMS 2000 subject classifications: Primary 62J99, 62H12; secondary 60G57.
Keywords and phrases: Varying coefficient model, low rank matrix estimation, statistical learning.

Received November 2012.

Contents

[*]Marianna Pensky was supported in part by National Science Foundation (NSF), grant DMS-1106564.

1. Introduction

In the present paper we consider the varying coefficient model which represents a useful tool for exploring dynamic patterns in economics, epidemiology, ecology, etc. This model can be viewed as a natural extension of the classical linear regression model and allows parameters that are constant in regression model to evolve with certain characteristics of the system such as time or age in epidemiological studies.

The varying coefficient models were introduced by Cleveland, Grosse and Shyu [4] and Hastie and Tibshirani [7] and have been extensively studied in the past 15 years. The estimation procedures for varying coefficient model are e.g. based on the kernel-local polynomial smoothing (see e.g. [28, 8, 5, 12]), the polynomial spline (see e.g. [9, 11, 10]), the smoothing spline (see e.g. [7, 8, 3]). More recently e.g. Wang et al [27] proposed a new procedure based on a local rank estimator; Kai et al [13] introduced a semi-parametric quantile regression procedure and studied an effective variable selection procedure; Lian [20] developed a penalization based approach for both variable selection and constant coefficient identification in a consistent framework. For more detailed discussions of the existing methods and possible applications, we refer to the very interesting survey of Fan and Zhang [6].

Existing methods typically provide asymptotic evaluation of precision of estimation procedures under the assumption that the number of observations tends to infinity. In practical applications, however, only a finite number of measurements are available. In the present paper, we focus on a non-asymptotic approach to the problem. We propose a novel estimation procedure which is based on recent developments in matrix estimation, in particular, matrix completion. In the matrix completion problem, one observes a small set of entries of a matrix and needs to estimate the remaining entries using these data. A standard assumption that allows such completion to be successful is that the unknown matrix has low rank or has approximately low rank. The matrix completion problem has attracted a considerable attention in the past few years (see, e.g., [2, 14, 19, 23, 16]). The most popular methods for matrix completion are based on nuclear-norm minimization which we adapt in the present paper.

1.1. Formulation of the problem

Let (W_i, t_i, Y_i), $i = 1, \dots, n$ be sampled independently from the varying coefficient model

$$Y = W^T f(t) + \sigma\xi. \tag{1}$$

Here, $W \in \mathbb{R}^p$ are random vectors of predictors, $f(\cdot) = (f_1(\cdot), \dots, f_p(\cdot))^T$ is an unknown vector-valued function of regression coefficients and $t \in [0,1]$ is a random variable independent of W. Let μ denote its distribution. The noise variable ξ is independent of W and t and is such that $\mathbb{E}(\xi) = 0$ and $\mathbb{E}(\xi^2) = 1$, $\sigma > 0$ denotes the noise level.

The goal is to estimate the vector function $f(\cdot)$ on the basis of observations (W_i, t_i, Y_i), $i = 1, \dots, n$. Our estimation method is based on the approximation of the unknown functions $f_i(t)$ using a basis expansion. This approximation generates the coordinate matrix A_0. In the above model, some of the components of vector function f are constant. The larger the part of the constant regression coefficients, the smaller the rank of the coordinate matrix A_0 (the rank of matrix A_0 does not exceed the number of time-varying components of vector $f(\cdot)$ by more than one). We suppose that the first element of this basis is just a constant function on $[0,1]$ (indeed, this is true for vast majority of bases on a finite interval). In this case, if the component $f_i(\cdot)$ is constant, then, it has only one non-zero coefficient in its expansion over the basis. This suggest the idea to take into account the number of constant regression coefficients using the rank of the coordinate matrix A_0.

Our procedure involves estimating A_0 using nuclear-norm penalization which is now a well-established proxy for rank penalization in the compressed sensing literature. Subsequently, the estimator of the coordinate matrix is plugged into the expansion yielding the estimator $\hat{f}(\cdot) = (\hat{f}_1(\cdot), \dots, \hat{f}_p(\cdot))^T$ of the vector function $f(t)$. For this estimator we obtain upper bounds on the mean squared error $\frac{1}{p}\sum\limits_{i=1}^{p} \|\hat{f}_i - f_i\|^2_{L_2(d\mu)}$ and on the pointwise estimation error $\frac{1}{p}\sum\limits_{i=1}^{p} |\hat{f}_i(t) - f_i(t)|$ for any $t \in \mathrm{supp}(\mu)$ (Corollary 1). These oracle inequalities are non-asymptotic and hold for finite values of p and n. The results in this paper concern random measurements and random noise and so they hold with high probability.

1.2. Layout of the paper

The remainder of this paper is organized as follows. In Section 1.3 we introduce notations used throughout the paper. In Section 2, we describe in details our estimation method, give examples of the possible choices of the basis (Section 2.1) and introduce an estimator for the coordinate matrix A_0 (Section 2.2). Section 3 presents the main results of the paper. In particular, Theorems 1 and 2 in Section 3 establish upper bounds for estimation error of the coordinate matrix A_0 measured in Frobenius norm. Corollary 1 provides non-asymptotic upper bounds for the mean squared and pointwise risks of the estimator of the vector function f. Section 4 considers an important particular case of the orthogonal dictionary.

1.3. Notations

We provide a brief summary of the notation used throughout this paper. Let A, B be matrices in $\mathbb{R}^{p\times l}$, μ be a probability distribution on $(0,1)$ and $\psi(\cdot)$ be a vector-valued function.

- For any vector $\eta \in \mathbb{R}^p$, we denote the standard l_1 and l_2 vector norms by $\|\eta\|_1$ and $\|\eta\|_2$, respectively.
- $\|\cdot\|_{L_2(d\mu)}$ and $\langle \cdot\,,\cdot\rangle_{L_2(d\mu)}$ are the norm and the scalar product in the space $L_2\left((0,1),d\mu\right)$.
- For $\psi(\cdot) = \left(\psi_1(\cdot),\dots,\psi_p(\cdot)\right)^T$, we set $\|\psi(\cdot)\|_\infty = \max\limits_{i=1,\dots,p}\ \sup\limits_{t\in\mathrm{supp}(\mu)} |\psi_i(t)|$ and $\|\psi(\cdot)\|_{L_2(d\mu)} = \max\limits_{1\le i\le p} \|\psi_i\|_{L_2(d\mu)}$
- We define the *scalar product of matrices* $\langle A,B\rangle = \mathrm{tr}(A^TB)$ where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix.
- Let
$$\|A\|_* = \sum_{j=1}^{\min(p,l)} \sigma_j(A) \quad \text{and} \quad \|A\|_2 = \left(\sum_{j=1}^{\min(p,l)} \sigma_j^2(A)\right)^{1/2}$$
be respectively the *trace* and *Frobenius* norms of the matrix A. Here $(\sigma_j(A))_j$ are the singular values of A ordered decreasingly.
- Let $\|A\| = \sigma_1(A)$.
- For any numbers, a and b, denote $a\vee b = \max(a,b)$ and $a\wedge b = \min(a,b)$.
- Denote the $k\times k$ identity matrix by $\mathbb{I}_k$.
- Let $(s-1)$ denote the number of non-constant $f_i(\cdot)$.
- 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.

2. Estimation method

The first step of our estimation method is the approximation of the unknown functions $f_i(t)$ by expanding them over an appropriate basis. This approximation generates the coordinate matrix A_0. Matrix A_0 is estimated using penalized risk minimization. The estimator of the coordinate matrix is plugged into the expansion yielding the estimator of the vector function f.

2.1. Basis expansion

Let $(\phi_i(\cdot))_{i=1,\dots,\infty}$ be an orthonormal basis in $L_2\left((0,1),d\mu\right)$, $l\in\mathbb{N}$ and $\phi(\cdot) = \left(\phi_1(\cdot),\dots,\phi_l(\cdot)\right)^T$. We assume that basis functions satisfy the following condition: there exists $c_\phi < \infty$ such that

$$\left\|\phi^T(t)\right\|_2^2 = \sum_{j=1}^{l} |\phi_j(t)|^2 \le c_\phi^2\, l, \tag{2}$$

for any $l \geq 1$ and any $t \in [0,1]$. Note that this condition is satisfied for most of the usual bases.

We introduce the coordinate matrix $A_0 \in \mathbb{R}^{p \times l}$ with elements

$$a^0_{kj} = \langle f_k, \phi_j \rangle_{L_2(d\mu)}, \quad k = 1, \dots, p,\ j = 1, \dots, l.$$

For each $k = 1, \dots, p$, we have

$$f_k(t) = \sum_{j=1}^{l} a^0_{kj} \phi_j(t) + \rho^{(l)}_k(t). \tag{3}$$

Denote the remainder by $\rho^{(l)}(\cdot) = (\rho^{(l)}_1(\cdot), \dots, \rho^{(l)}_p(\cdot))^T$. We assume that the basis $(\phi_i(\cdot))_{i=1,\dots,\infty}$ guarantees good approximation of f_k by $\sum_{j=1}^{l} a^0_{kj} \phi_j(t)$, that is,

Assumption 1. *We assume that the basis satisfies condition (2) and that there exists a positive constant b such that, for any $l \geq 1$*

$$\left\| \rho^{(l)}(\cdot) \right\|_{\infty} \quad \leq \quad b\, l^{-\gamma}, \quad \gamma > 0. \tag{4}$$

Often approximation in L_2-norm gives better rates of convergence. In order to get upper bounds on the mean squared error we will use the following additional assumption:

Assumption 2. *There exist $b_1 > 0$ such that, for any $l \geq 1$*

$$\left\| \rho^{(l)}(\cdot) \right\|_{L_2(d\mu)} \leq b_1\, l^{-(\gamma+1/2)}, \quad \gamma > 0.$$

Let us give few examples of possible choices of the basis.

Example 1. Assume that $d\mu = g(t)\,dt$ and function g is bounded away from zero and infinity, i.e. there exist absolute constants g_1 and g_2 such that for any $t \in \mathrm{supp}(\mu)$

$$g_1 \leq g(t) \leq g_2, \quad 0 < g_1 < g_2 < \infty. \tag{5}$$

Denote $\widetilde{\phi}_j(t) = e^{2\,i\,\pi\,j\,t}$, $j \in \mathbb{Z}$, the standard Fourier basis of $L_2((0,1))$. Then, it is easy to check that $\phi_j(t) = \widetilde{\phi}_j(t)/\sqrt{g(t)}$, $j \in \mathbb{Z}$, is an orthonormal basis of $L_2((0,1), g)$. Moreover, condition (2) holds with $c^2_\phi = g_1^{-1}$.

For $\gamma > 0$, consider the Sobolev space $\mathbb{W}_\gamma(0,1)$ of functions $F \in L_2(0,1)$ with the norm $\|F\|^2_{\mathbb{W}_\gamma} = \int_{-\infty}^{\infty} |\omega|^{2\gamma+1} |\hat{F}(\omega)|^2 d\omega$ where $\hat{F}(\omega)$ is the Fourier transform of F. Then, by Theorems 9.1 and 9.2 of [22], one has

$$\sum_{j=-\infty}^{\infty} |j|^{2\gamma+1} |\langle F, \widetilde{\phi}_j \rangle|^2 \leq C_\gamma \|F\|^2_{\mathbb{W}_\gamma}, \tag{6}$$

where C_γ is an absolute constant which depends on γ only. Assume that for some $A < \infty$ the functions f_k belong to a Sobolev ball of radius A, i.e.

$$\max_{k=1,\dots,p} \left\| f_k(\cdot) \sqrt{g(\cdot)} \right\|_{\mathbb{W}_\gamma} \leq A, \quad \gamma > 0. \tag{7}$$

Let $l = 2N + 1$, so that

$$f_k(t) = \sum_{j=-N}^{N} a^0_{kj} \phi_j(t), \quad \rho^{(l)}_k(t) = \sum_{|j|>N} a^0_{kj} \phi_j(t),$$

where $a^0_{kj} = \langle f_k(t)\sqrt{g(t)}, \widetilde{\phi}_j(t)\rangle$. Then, it follows from equations (5), (6) and (7) that

$$\begin{aligned}\left\|\rho^{(l)}(\cdot)\right\|^2_\infty &\le g_1^{-1} \left[\sum_{|j|>N} |j|^{-2\gamma-1}\right] \left[\max_{k=1,\ldots,p} \sum_{|j|>N} |j|^{2\gamma+1} |a^0_{kj}|^2\right] \\ &\le \frac{A^2 C_\gamma}{g_1} \sum_{|j|>N} |j|^{-2\gamma-1} \le \frac{A^2 C_\gamma}{2\, g_1\, \gamma\, N^{2\gamma}}\end{aligned}$$

where $N = (l-1)/2$ and

$$\left\|\rho^{(l)}(\cdot)\right\|^2_{L_2(g)} \le N^{-(2\gamma+1)} A^2 C_\gamma,$$

so that Assumptions 1 and 2 hold.

Example 2. Consider a wavelet ψ with a bounded support of length C_ψ and with γ^* vanishing moments and choose $l = 2^H$ where H is a positive integer. Construct a periodic wavelet basis $\psi_{h,i}(t)$, $h = -1, \ldots, J-1$, $i = 0, \ldots, 2^h - 1$, with $\psi_{-1,0}(t) = 1$ and $\psi_{h,i}(t) = 2^{h/2}\psi(2^h t - i)$ for $h \ge 0$. As in Example 1, set $\phi_j(t) = \phi_{h,i}(t) = \psi_{h,i}(t)/\sqrt{g(t)}$ where $j = 2^h + i + 1$. Note that condition (2) holds in this case with $c^2_\phi = g_1^{-1} C_\psi \|\psi\|^2_\infty$.

Then, each function $f_k(t)$ can be expanded into a wavelet series

$$f_k(t) = \sum_{h=-1}^{H-1} \sum_{i=0}^{2^h-1} a^0_{k,h,i}\, \phi_{h,i}(t), \quad \rho^{(l)}_k(t) = \sum_{h=H}^{\infty} \sum_{i=0}^{2^h-1} a^0_{k,h,i}\, \phi_{h,i}(t),$$

where $a^0_{k,h,i} = \langle f_k(\cdot)\sqrt{g(\cdot)}, \psi_{h,i}(\cdot)\rangle$.

Theorem 9.4 of [22] states that for $F \in \mathbb{W}_\gamma(0,1)$ one has

$$\sum_{h=-1}^{\infty} 2^{h(2\gamma+1)} \sum_{i=0}^{2^h-1} |\langle F, \psi_{h,i}\rangle|^2 \le C_\gamma \|F\|^2_{\mathbb{W}_\gamma},$$

where C_γ is an absolute constant which depends on γ only, provided $\gamma < \gamma^*$. Then, under assumptions (5) and (7), as in Example 1, Assumption 1 holds. For example, recalling that $H = \log_2 l$ and that length of support of ψ is bounded by C_ψ, obtain

$$\left\|\rho^{(l)}(\cdot)\right\|^2_{L_2(g)} \le 2^{-H(2\gamma+1)} \max_{k=1,\ldots,p} \sum_{h=H}^{\infty} 2^{h(2\gamma+1)} \sum_{i=0}^{2^h-1} |a^0_{k,h,i}|^2 \le A^2 C_\gamma l^{-(2\gamma+1)},$$

$$\begin{aligned}\left\|\rho^{(l)}(\cdot)\right\|_\infty^2 &\leq (2\gamma g_1)^{-1} C_\psi \left\|\psi\right\|_\infty^2 2^{-2H\gamma} \max_{k=1,\dots,p} \sum_{h=-1}^{\infty} 2^{h(2\gamma+1)} \sum_{i=0}^{2^h-1} |a^0_{k,h,i}|^2 \\ &\leq A^2 (2\gamma g_1)^{-1} C_\psi \left\|\psi\right\|_\infty^2 l^{-2\gamma},\end{aligned}$$

where $\|\psi\|_\infty = \sup_t |\psi(t)|$.

Example 3. Suppose that $f_i(t)$ belong to a finite $k-$dimensional sub-space of $L_2\left((0,1), d\mu\right)$. For example, $f_i(t)$ are polynomials of degree less than k. Then, choosing $l = k$ and an orthonormal basis in this sub-space, we have trivially $\rho^{(l)}(\cdot) = 0$.

2.2. Estimation of the coordinate matrix

Denoting $X = W\phi^T(t)$, we can rewrite (1) in the following form

$$Y = \operatorname{tr}\left(A_0 X^T\right) + W^T \rho^{(l)}(t) + \sigma\xi. \tag{8}$$

We suppose that some of the functions $f_i(\cdot)$ are constant and let $(s-1)$ denote the number of non-constant $f_i(\cdot)$. This parameter, s, plays an important role in what follows. Note that $\operatorname{rank}(A_0) \leq s$.

Using observations (Y_i, X_i) we define the following estimator of A_0:

$$\hat{A} = \arg\min \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle X_i, A\rangle\right)^2 + \lambda \left\|A\right\|_* \right\}, \tag{9}$$

where λ is the regularization parameter. This penalization, using the trace-norm, is now quite standard in matrix completion problem and allows one to recover a matrix from under-sampled measurements.

Using estimator (9) of the coordinate matrix A_0, we recover $f(t)$ as

$$\hat{f}(t) = \hat{A}\phi(t).$$

2.3. Assumptions about the dictionary and the noise

We assume that the vectors W_i are i.i.d copies of a random vector W having distribution Π on a given set of vectors $\mathcal{X}$. Using rescaling, we can suppose that $\left\|W\right\|_2 \leq 1$ almost surely. Let $\mathbb{E}\left(W\,W^T\right) = \Omega$ and $\omega_{\max}, \omega_{\min}$ denote respectively its maximal and minimal singular values. We need the following assumption on the distribution of W.

Assumption 3. *The matrix $\Omega = \mathbb{E}\left(W\,W^T\right)$ is positive definite.*

Let $\|A\|^2_{L_2(\Pi\otimes\mu)} = \mathbb{E}\left(\langle X, A\rangle^2\right)$. An easy computation leads to

$$\begin{aligned}\|A\|^2_{L_2(\Pi\otimes\mu)} &= \mathbb{E}\left(\langle W, A\,\phi(t)\rangle^2\right) \\ &= \mathbb{E}_t\left(\mathbb{E}_W\left(\langle W, A\,\phi(t)\rangle^2\right)\right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_W\left(\langle W, A\,\phi(t)\rangle^2\right) &= \mathbb{E}_W\left(\operatorname{tr}\left((A\,\phi(t))^T\,W\,W^T A\,\phi(t)\right)\right)\\ &= \mathbb{E}_W\left(\operatorname{tr}\left(W\,W^T A\,\phi(t)\,(A\,\phi(t))^T\right)\right)\\ &= \left\langle \mathbb{E}_W\left(W^T\,W\right), A\,\phi(t)\,(A\,\phi(t))^T\right\rangle\\ &= \left\langle \Omega, A\,\phi(t)\,(A\,\phi(t))^T\right\rangle.\end{aligned}$$

By definition we obtain

$$\left\langle \Omega, A\,\phi(t)\,(A\,\phi(t))^T\right\rangle \geq \omega_{\min}\,\|A\,\phi(t)\|_2^2\,.$$

Finally we compute

$$\|A\|^2_{L_2(\Pi\otimes\mu)} \geq \omega_{\min}\,\mathbb{E}_t\left(\|A\,\phi(t)\|_2^2\right) = \omega_{\min}\,\|A\|_2^2 \tag{10}$$

where in the last display we used that $(\phi_i(\cdot))_{i=1,\dots,\infty}$ is an orthonormal basis in $L_2\left((0,1), d\mu\right)$.

We consider the case of *sub-exponential noise* which satisfies the following condition

Assumption 4. *There exist a constant $K > 0$ such that*

$$\max_{i=1,\dots,n}\mathbb{E}\exp\left(|\xi_i|/K\right) \leq e.$$

For instance, if ξ_i are i.i.d. standard Gaussian we can take $K = 1$.

3. Main Results

Let

$$\Sigma_R = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_i \qquad \text{and} \qquad \Sigma = \frac{1}{n}\sum_{i=1}^{n}\left(W_i^T\rho^{(l)}(t_i) + \sigma\,\xi_i\right)X_i$$

where $\{\epsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence. These stochastic terms play an important role in the choice of the regularization parameter λ.

We introduce the following notations:

$$M = \operatorname{tr}(\Omega)\vee(l\,\omega_{\max}) \quad \text{and} \quad n^{**} = \frac{C\,c_\phi^2\,l\,\log(d)}{\omega_{\min}^2}\left[(M\,s)\vee 1\right].$$

The following theorem gives a general upper bound on the prediction error for the estimator $\hat{A}$ given by (9). Its proof is given in Appendix A.

Theorem 1. *Let $\lambda \geq 3\,\|\Sigma\|$ and suppose that Assumption 3 holds. Then, with probability at least $1 - 2/d$,*

(i)

$$\left\|\hat{A}-A_0\right\|_2^2 \leq C\,\max\left\{\frac{s}{\omega_{\min}^2}\left(\lambda^2+\|A_0\|_*^2\,\frac{c_\phi^2\,l\,M\,\log(d)}{n}\right),\,\frac{c_\phi\,\|A_0\|_*^2}{\omega_{\min}}\sqrt{\frac{\log(d)\,l}{n}}\right\}.$$

(ii) If, in addition $n \geq n^{**}$, then

$$\left\|\hat{A}-A_0\right\|_2^2 \leq \frac{C\,s\,\lambda^2}{\omega_{\min}^2}$$

where $d = l + p$.

In order to obtain upper bounds in Theorem 1 in a closed form, it is necessary to obtain a suitable upper bound for $\|\Sigma\|$. The following lemma, proved in Section E, gives such bound.

Lemma 1. *Under Assumptions 1 - 4, there exists a numerical constant c^*, that depends only on K, such that, for all $t > 0$ with probability at least $1 - 2e^{-t}$*

$$\|\Sigma\| \leq \left(\sigma\,c^* + \frac{2\,b\sqrt{s-1}}{l^\gamma}\right)\max\left\{\sqrt{\frac{M\,(t+\log(d))}{n}},\ \frac{c_\phi\,\sqrt{l}\,\,(t+\log(d))\left(\left[K\,\log\left(\frac{K\,c_\phi}{\omega_{\max}}\right)\right]\vee 1\right)}{n}\right\} \tag{11}$$

where $d = p + l$.

The optimal choice of the parameter t in Lemma 1 is $t = \log(d)$. Larger t leads to a slower rate of convergence and a smaller t does not improve the rate but makes the concentration probability smaller. With this choice of t, the second terms in the maximum in (11) is negligibly small for $n \geq n^*$ where

$$n^* = \frac{2\,c_\phi^2\,l\,\left(\left[K\,\log\left(\frac{K\,c_\phi}{\omega_{\max}}\right)\right]\vee 1\right)^2\,\log(d)}{M}.$$

In order to satisfy condition $\lambda \geq 3\,\|\Sigma\|$ in Theorem 1 we can choose

$$\lambda = 4.25\left(c^*\sigma + \frac{2\,b\sqrt{s-1}}{l^\gamma}\right)\sqrt{\frac{M\,\log(d)}{n}}. \tag{12}$$

If ξ_i are $N(0,1)$, then we can take $c^* = 6.5$ (see Lemma 4 in [15]).

With these choices of λ, we obtain the following theorem.

Theorem 2. *Let Assumptions 1 - 4 hold. Consider regularization parameters λ satisfying (12) and $n \geq n^*$. Then, with probability greater than $1 - 4/d$*

(i)

$$\left\|\hat{A}-A_0\right\|_2^2 \leq C\,\max\left\{\left(\sigma^2+\frac{b^2\,(s-1)}{l^{2\gamma}}+l\,\|A_0\|_*^2\right)\frac{M\,s\,\log(d)}{n\,\omega_{\min}^2},\right.$$
$$\left.\frac{c_\phi\,\|A_0\|_*^2}{\omega_{\min}}\sqrt{\frac{\log(d)\,l}{n}}\right\}.$$

(ii) If, in addition $n \geq n^{**}$, then

$$\left\|\hat{A}-A_0\right\|_2^2 \leq C\,\left(\sigma^2+\frac{b^2\,(s-1)}{l^{2\gamma}}\right)\frac{M\,s\,\log(d)}{n\,\omega_{\min}^2}.$$

Using $\hat{A}$ we define the estimator of $f(t)$ as

$$\hat{f}(t)=\left(\hat{f}_1(t),\ldots,\hat{f}_p(t)\right)^T=\hat{A}\,\phi(t). \tag{13}$$

Theorem 2 allows to obtain the following upper bounds on the prediction error of $\hat{f}(t)$.

Corollary 1. *Suppose that the assumptions of Theorem 2 hold. With probability greater than $1-4/d$, one has*

(a) $\forall t \in \operatorname{supp}(\mu)$

$$\frac{1}{p}\sum_{i=1}^{p}|\hat{f}_i(t)-f_i(t)| \leq \frac{C\,\|\phi(t)\|_2^2\,\beta}{n}+\frac{2\,b^2\,s}{p\,l^{2\gamma}},$$

(b) If, in addition, Assumption 2 holds

$$\frac{1}{p}\sum_{i=1}^{p}\|\hat{f}_i-f_i\|_{L_2(d\mu)}^2 \leq \frac{C\,\beta}{n}+\frac{2\,b_1^2\,s}{p\,l^{(2\gamma+1)}},$$

where

$$\beta=\begin{cases}\left(\sigma^2+\dfrac{b^2\,(s-1)}{l^{2\gamma}}\right)\dfrac{M\,s\,\log(d)}{p\,\omega_{\min}^2}, & \textit{if}\quad n\geq n^{**}\\[2ex] \max\left\{\left(\sigma^2+\dfrac{b^2(s-1)}{l^{2\gamma}}+l\|A_0\|_*^2\right)\dfrac{M\,s\,\log(d)}{p\,\omega_{\min}^2},\dfrac{c_\phi\|A_0\|_*^2\sqrt{\log(d)\,l\,n}}{\omega_{\min}\,p}\right\}, & \textit{if not.}\end{cases}$$

Proof. We shall prove the second statement of the corollary, the first one can be proved in a similar way. Let A^i denote the i-th row of a matrix A. We compute

$$\begin{aligned}\left\|f_i(t)-\hat{A}^i\phi(t)\right\|_{L_2(d\mu)} &\leq \left\|f_i(t)-A_0^i\phi(t)\right\|_{L_2(d\mu)}+\left\|\left(A_0^i-\hat{A}^i\right)\phi(t)\right\|_{L_2(d\mu)}\\ &=\left\|\rho_i^{(l)}(t)\right\|_{L_2(d\mu)}+\left\|A_0^i-\hat{A}^i\right\|_2\end{aligned} \tag{14}$$

where in the last display we used that $(\phi_i(\cdot))_{i=1,\dots,\infty}$ is an orthonormal basis. Using (14) and Assumption 2 we derive

$$\sum_{i=1}^{p} \|\hat{f}_i - f_i\|^2_{L_2(d\mu)} \leq \frac{2\, b_1^2\, s}{l^{(2\gamma+1)}} + 2\left\| \hat{A} - A_0 \right\|_2^2 .$$

Now Theorem 2 implies the statement of the corollary. □

4. Orthonormal dictionary

As an important particular case, let us consider the orthonormal dictionary. Let $(e_j)_j$ be the canonical basis of $\mathbb{R}^p$. Assume that the vectors W_i are i.i.d copies of a random vector W which has the uniform distribution Π on the set

$$\mathcal{X} = \{e_j,\ 1 \leq j \leq p\}.$$

Note that this is an unfavorable case of very "sparse observations", that is, each observation provides some information on only one of the coefficients of $f(t)$.

In this case, $\Omega = \frac{1}{p}\mathbb{I}_p$, $\omega_{\max} = \omega_{\min} = \frac{1}{p}$ and we obtain the following values of parameters

$$\begin{aligned} M &= \frac{l \vee p}{p}, \\ n^* &= 2\, K^2\, \log^2(K\, p)\; c_\phi^2\, \log(d)\,(l \wedge p), \\ \lambda &= 4.25\left(C^*\sigma + \frac{2\, b\sqrt{s-1}}{l^\gamma}\right)\sqrt{\frac{(l \vee p)\, \log(d)}{p\, n}}, \\ n^{**} &= C\, c_\phi^2\, l\, s\, p\,(l \vee p)\, \log(d). \end{aligned} \tag{15}$$

Plugging these values into Corollary 1, we derive the following result.

Corollary 2. *Let Assumptions 1 and 4 hold. Consider regularization parameter λ satisfying (15), and $n \geq n^*$. Then, with probability greater than $1 - 4/d$, one has*

(a) $\forall t \in \operatorname{supp}(\mu)$

$$\frac{1}{p}\sum_{i=1}^{p} |\hat{f}_i(t) - f_i(t)| \leq \frac{C\, \|\phi(t)\|_2^2\, \beta}{n} + \frac{2\, b^2\, s}{p\, l^{2\gamma}}, \tag{16}$$

(b) If, in addition, Assumption 2 holds

$$\frac{1}{p}\sum_{i=1}^{p} \|\hat{f}_i - f_i\|^2_{L_2(d\mu)} \leq \frac{C\, \beta}{n} + \frac{2\, b_1^2\, s}{p\, l^{(2\gamma+1)}}, \tag{17}$$

where

$$\beta = \begin{cases} \left(\sigma^2 + \dfrac{b^2\,(s-1)}{l^{2\gamma}}\right)(l \vee p)\, s\, \log(d), & \textit{if} \quad n \geq n^{**} \\[2ex] \left(\sigma^2 + \dfrac{b^2\,(s-1)}{l^{2\gamma}} + l\, \|A_0\|_*^2\right)(l \vee p)\, s\, \log(d), & \textit{if not.} \end{cases}$$

Remarks. *Optimal choice of parameter l:* The upper bounds given in Corollary 2 indicate the optimal choice of parameter l. From (15) we compute the following values of l:

$$l_1^* = \frac{n}{C\, c_\phi^2\, s\, p^2\, \log(d)} \quad \text{if} \quad l \le p$$

and

$$l_2^* = \sqrt{\frac{n}{C\, c_\phi^2\, s\, p\, \log(d)}} \quad \text{if} \quad l > p.$$

Let

$$\begin{aligned}
F_1(l) &= C\left(\sigma^2 + \frac{b^2\,(s-1)}{l^{2\gamma}}\right)\frac{p\, s\, \log(d)}{n} + \frac{2\, b_1^2\, s}{p\, l^{(2\gamma+1)}},\\
F_2(l) &= F_1(l) + l\, \|A_0\|_*^2\, \frac{p\, s\, \log(d)}{n},\\
F_3(l) &= C\left(\sigma^2 + \frac{b^2\,(s-1)}{l^{2\gamma}}\right)\frac{l\, s\, \log(d)}{n} + \frac{2\, b_1^2\, s}{p\, l^{(2\gamma+1)}},\\
F_4(l) &= F_3(l) + l^2\, \|A_0\|_*^2\, \frac{s\, \log(d)}{n}.
\end{aligned}$$

Let $\gamma \ge 1/2$ and consider first the case $s\, p^3\, \log(d) \gtrsim n \gtrsim s\, p^2\, \log(d)$ (the symbol $\lesssim$ means that the inequality holds up to a multiplicative numerical constant). Then, Corollary 2 implies that

$$\frac{1}{p}\sum_{i=1}^{p}\|\hat{f}_i - f_i\|_{L_2(d\mu)}^2 \le \begin{cases} F_1(l), & \text{if} \quad 1 \le l \le l_1^* \\ F_2(l), & \text{if} \quad l_1^* < l \le p \\ F_4(l), & \text{if} \quad l > p. \end{cases}$$

On $[1, l_1^*]$, $F_1(l)$ achieves its minimum at l_1^*. Note that $F_1(l_1^*) \le F_2(l)$ for any $l \in [l_1^*, p]$ and $F_1(l_1^*) \le F_4(l)$ for any $l > p$. Then, for $s\, p^3\, \log(d) \gtrsim n \gtrsim s\, p^2\, \log(d)$ the optimal value of l minimizing (17) is

$$\hat{l}_1 = \left[\frac{n}{C\, c_\phi^2\, s\, p^2\, \log(d)}\right].$$

When $n \gtrsim s\, p^3\, \log(d)$, the Corollary 2 implies that

$$\frac{1}{p}\sum_{i=1}^{p}\|\hat{f}_i - f_i\|_{L_2(d\mu)}^2 \le \begin{cases} F_1(l), & \text{if} \quad 1 \le l \le p \\ F_3(l), & \text{if} \quad p < l \le l_2^* \\ F_4(l), & \text{if} \quad l > l_2^*. \end{cases}$$

Let

$$l_3^* = \left(\frac{C\, n}{\sigma^2\, p\, \log(d)}\right)^{\frac{1}{2\gamma+2}}.$$

On $[p, l_2^*]$, $F_3(l)$ achieves its minimum at l_2^* if $p^{3+2\gamma}\log(d) \gtrsim n \gtrsim s\,p^3\log(d)$ and at l_3^* if $n \gtrsim p^{3+2\gamma}\log(d)$. Note that $F_3(l_2^*) \leq F_1(l)$ for any $l \in [1,p]$ and $F_3(l_2^*) \leq F_4(l)$ for any $l > l_2^*$. Then, for $p^{3+2\gamma}\log(d) \gtrsim n \gtrsim s\,p^3\log(d)$ the optimal value of l minimizing (17) is

$$\hat{l}_2 = \left\lceil \sqrt{\frac{n}{C\,c_\phi^2\,s\,p\,\log(d)}}\right\rceil$$

and for $n \gtrsim p^{3+2\gamma}\log(d)$ the optimal value of l is

$$\hat{l}_3 = \left(\frac{C\,n}{\sigma^2\,p\,\log(d)}\right)^{\frac{1}{2\gamma+2}}.$$

Minimax rate of convergence: For $p = 1$ the optimal choice of l in (17) is

$$\hat{l} = \left(\frac{2\,(2\,\gamma+1)\,b^2\,n}{\sigma^2\,\log(d)}\right)^{\frac{1}{2\gamma+2}}.$$

With this choice of l, the rate of convergence given by Corollary 2 is $n^{-\frac{2\gamma+1}{2\gamma+2}}$. Note that for $f \in \mathbb{W}_\gamma(0,1)$ we recover the minimax rate of convergence as given in e.g. [26].

Acknowledgements

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

Appendix

Appendix A: Proof of Theorem 1

This proof uses ideas developed in the proof of Theorem 3 in [16]. The main difference is that here we have no restriction on the sup −norm of A_0. This implies several modifications in the proof.

It follows from the definition of the estimator $\hat{A}$ that

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left\langle X_i, \hat{A}\right\rangle\right)^2 + \lambda\|\hat{A}\|_* \leq \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle X_i, A_0\rangle\right)^2 + \lambda\|A_0\|_*$$

which, due to (8), implies

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left\langle X_i, A_0 - \hat{A}\right\rangle + W_i^T\rho^{(l)}(t_i) + \xi_i\right)^2 + \lambda\|\hat{A}\|_* \leq \frac{1}{n}\sum_{i=1}^{n}\left(W_i^T\rho^{(l)}(t_i) + \xi_i\right)^2 + \lambda\|A_0\|_*. \tag{18}$$

Set $H = A_0 - \hat{A}$ and $\Sigma = \frac{1}{n}\sum\limits_{i=1}^{n}\left(W_i^T\rho^{(l)}(t_i) + \xi_i\right)X_i$. Then, we can write (18) in the following way

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i, H\rangle^2 + 2\langle \Sigma, H\rangle + \lambda\|\hat{A}\|_* \leq \lambda\|A_0\|_*.$$

By duality between the nuclear and the operator norms, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i, H\rangle^2 + \lambda\|\hat{A}\|_* \leq 2\|\Sigma\|\,\|H\|_* + \lambda\|A_0\|_*. \tag{19}$$

Let P_S denote the projector on the linear subspace S and let $S^\perp$ be the orthogonal complement of S. Let $u_j(A)$ and $v_j(A)$ denote respectively the *left* and the *right* orthonormal *singular vectors* of A, $S_1(A)$ is the linear span of $\{u_j(A)\}$, $S_2(A)$ is the linear span of $\{v_j(A)\}$. For $A, B \in \mathbb{R}^{p\times l}$ we set $\mathbf{P}_A^\perp(B) = P_{S_1^\perp(A)}BP_{S_2^\perp(A)}$ and $\mathbf{P}_A(B) = B - \mathbf{P}_A^\perp(B)$.

By definition, for any matrix B, the singular vectors of $\mathbf{P}_{A_0}^\perp(B)$ are orthogonal to the space spanned by the singular vectors of A_0. This implies that $\left\|A_0 + \mathbf{P}_{A_0}^\perp(H)\right\|_1 = \|A_0\|_* + \left\|\mathbf{P}_{A_0}^\perp(H)\right\|_*$. Then we compute

$$\begin{aligned}\left\|\hat{A}\right\|_* &= \left\|A_0 + H\right\|_*\\ &= \left\|A_0 + \mathbf{P}_{A_0}^\perp(H) + \mathbf{P}_{A_0}(H)\right\|_*\\ &\geq \left\|A_0 + \mathbf{P}_{A_0}^\perp(H)\right\|_* - \|\mathbf{P}_{A_0}(H)\|_*\\ &= \|A_0\|_* + \left\|\mathbf{P}_{A_0}^\perp(H)\right\|_* - \|\mathbf{P}_{A_0}(H)\|_*.\end{aligned} \tag{20}$$

From (20) we obtain

$$\|A_0\|_* - \left\|\hat{A}\right\|_* \leq \|\mathbf{P}_{A_0}(H)\|_* - \left\|\mathbf{P}_{A_0}^\perp(H)\right\|_*. \tag{21}$$

From (19), using (21) and $\lambda \geq 3\|\Sigma\|$ we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\langle X_i, H\rangle^2 &\leq 2\|\Sigma\|\,\|\mathbf{P}_{A_0}(H)\|_* + \lambda\,\|\mathbf{P}_{A_0}(H)\|_*\\ &\leq \frac{5}{3}\lambda\,\|\mathbf{P}_{A_0}(H)\|_*.\end{aligned} \tag{22}$$

Since $\mathbf{P}_A(B) = P_{S_1^\perp(A)}BP_{S_2(A)} + P_{S_1(A)}B$ and $\operatorname{rank}(P_{S_i(A)}B) \leq \operatorname{rank}(A)$ we derive that $\operatorname{rank}(\mathbf{P}_A(B)) \leq 2\operatorname{rank}(A)$. From (22) we compute

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i, H\rangle^2 \leq \frac{5}{3}\lambda\sqrt{2R}\,\|H\|_2 \tag{23}$$

where we set $R = \operatorname{rank}(A_0)$.

For $0 < r \leq m = \min(p, l)$ we consider the following constraint set

$$\mathcal{C}(r) = \left\{ \|A\|_2 \leq 1,\ \|A\|^2_{L_2(\Pi\otimes\mu)} \geq c_\phi \sqrt{\frac{64\ \log(d)\, l}{\log(6/5)\ n}}, \|A\|_* \leq \sqrt{r}\, \|A\|_2 \right\} \tag{24}$$

where $\|A\|^2_{L_2(\Pi\otimes\mu)} = \mathbb{E}\left(\langle X, A\rangle^2\right)$. Note that the condition $\|A\|_* \leq \sqrt{r}\,\|A\|_2$ is satisfied if $\operatorname{rank}(A) \leq r$.

The following lemma shows that for matrices $A \in \mathcal{C}(r)$ we have some approximative restricted isometry. Its proof is given in Appendix B.

Lemma 2. *For all $A \in \mathcal{C}(r)$*

$$\frac{1}{n}\sum_{i=1}^{n} \langle X_i, A\rangle^2 \geq \frac{1}{2}\|A\|^2_{L_2(\Pi\otimes\mu)} - \frac{44\, c_\phi^2\, l\, r}{\omega_{\min}} \left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2$$

with probability at least $1 - \frac{2}{d}$.

We need the following auxiliary lemma which is proved in Appendix D.

Lemma 3. *If $\lambda_1 > 3\,\|\Sigma\|$*

$$\left\|\mathbf{P}^{\perp}_{A_0}(H)\right\|_* \leq 5\,\|\mathbf{P}_{A_0}(H)\|_*\,.$$

Lemma 3 implies that

$$\begin{aligned}\|H\|_* &\leq 6\,\|\mathbf{P}_{A_0}(H)\|_* \\ &\leq \sqrt{72\,R}\,\|H\|_2\,.\end{aligned} \tag{25}$$

If $\|H\|^2_{L_2(\Pi\otimes\mu)} \geq c_\phi\ \|H\|_2^2\ \sqrt{\frac{64\ \log(d)\, l}{\log(6/5)\, n}}$, (25) implies that $\frac{H}{\|H\|_2} \in \mathcal{C}\,(72\,R)$ and we can apply Lemma 2. From Lemma 2 and (23) we obtain that with probability at least $1 - \frac{2}{d}$ one has

$$\frac{1}{2}\|H\|^2_{L_2(\Pi\otimes\mu)} \leq \frac{5}{3}\lambda\sqrt{2\,R}\,\|H\|_2 + \frac{3168\, c_\phi^2\, l\, R}{\omega_{\min}}\,\|H\|_2^2\,\left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2. \tag{26}$$

The following Lemma, proved in Section E.2, gives a suitable bound on $\mathbb{E}\,\|\Sigma_R\|$:

Lemma 4. *Let $(\epsilon_i)_{i=1}^n$ be an i.i.d. Rademacher sequence. Suppose that Assumption 3 holds. Then,*

$$\mathbb{E}\,\|\Sigma_R\| \leq 4.6\sqrt{\frac{M\ \log(d)}{n}}$$

where $d = p + l$ and $M = \operatorname{tr}(\Omega) \vee (l\omega_{\max})$.

Using Lemma 4, (10) and (26) we obtain

$$\omega_{\min}\|H\|_2^2 \leq \frac{10}{3}\lambda\sqrt{2\,R}\,\|H\|_2 + \frac{C\, c_\phi^2\, l\, R\, M\ \log(d)}{\omega_{\min}\, n}\,\|H\|_2^2\,. \tag{27}$$

On the other hand, equation (19) and the triangle inequality imply that

$$\lambda\|\hat{A}\|_* \leq 2\,\|\Sigma\|\,\|\hat{A}\|_* + 2\,\|\Sigma\|\,\|A_0\|_* + \lambda\|A_0\|_*$$

and $\lambda \geq 3\,\|\Sigma\|$ gets

$$\|\hat{A}\|_2 \leq \|\hat{A}\|_* \leq 5\|A_0\|_*. \tag{28}$$

Putting (28) into (27) and using $\operatorname{rank}(A_0) \leq s$ we compute

$$\|H\|_2^2 \leq \frac{C\,s}{\omega_{\min}^2}\left(\lambda^2 + \frac{c_\phi^2\, l\, M\, \log(d)\, \|A_0\|_*^2}{n}\right)$$

which implies the statement (i) of Theorem 1 in the case when $\|H\|_{L_2(\Pi\otimes\mu)}^2 \geq c_\phi\,\|H\|_2^2\,\sqrt{\frac{64\,\log(d)\,l}{\log(6/5)\,n}}$.

If $\|H\|_{L_2(\Pi\otimes\mu)}^2 \leq c_\phi\,\|H\|_2^2\,\sqrt{\frac{64\,\log(d)\,l}{\log(6/5)\,n}}$, using (10), we derive

$$\omega_{\min}\,\|H\|_2^2 \leq c_\phi\,\|H\|_2^2\,\sqrt{\frac{64\,\log(d)\,l}{\log(6/5)\,n}}. \tag{29}$$

Then (28) implies

$$\|H\|_2^2 < \frac{C\,c_\phi\,\|A_0\|_*^2}{\omega_{\min}}\sqrt{\frac{\log(d)\,l}{n}}.$$

This completes the proof of part (i) of Theorem 1.

If, in addition $n > 2\,\frac{C\,c_\phi^2\,l\,s\,M\,\log(d)}{\omega_{\min}^2}$, from (27) we obtain

$$\omega_{\min}\|H\|_2^2 \leq \frac{10}{3}\lambda\sqrt{2\,R}\,\|H\|_2 + \frac{\omega_{\min}}{2}\,\|H\|_2^2$$

and

$$\|H\|_2^2 \leq \frac{C\,s\,\lambda^2}{\omega_{\min}^2}.$$

On the other hand, for $n > n^{**}$ (29) does not hold. This completes the proof of Theorem 1.

Appendix B: Proof of Lemma 2

Set $\mathcal{E} = \frac{44\,c_\phi^2\,l\,r(\mathbb{E}(\|\Sigma_R\|))^2}{\omega_{\min}}$. We will show that the probability of the following bad event is small

$$\mathcal{B} = \left\{\exists\, A \in \mathcal{C}(r)\,\text{such that}\,\left|\frac{1}{n}\sum_{i=1}^{n}\langle X_i, A\rangle^2 - \|A\|_{L_2(\Pi\otimes\mu)}^2\right| > \frac{1}{2}\|A\|_{L_2(\Pi\otimes\mu)}^2 + \mathcal{E}\right\}.$$

Note that $\mathcal{B}$ contains the complement of the event that we are interested in.

In order to estimate the probability of $\mathcal{B}$ we use a standard peeling argument. Let $\nu = c_\phi \sqrt{\frac{64\,\log(d)\,l}{\log(6/5)\,n}}$ and $\alpha = \frac{6}{5}$. For $k \in \mathbb{N}$ set

$$S_k = \left\{ A \in \mathcal{C}(r)\,:\, \alpha^{k-1}\nu \leq \|A\|^2_{L_2(\Pi\otimes\mu)} \leq \alpha^k \nu \right\}.$$

If the event $\mathcal{B}$ holds for some matrix $A \in \mathcal{C}(r)$, then A belongs to some S_k and

$$\begin{aligned} \left| \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2 - \|A\|^2_{L_2(\Pi\otimes\mu)} \right| &> \frac{1}{2}\|A\|^2_{L_2(\Pi\otimes\mu)} + \mathcal{E} \\ &> \frac{1}{2}\alpha^{k-1}\nu + \mathcal{E} \\ &= \frac{5}{12}\alpha^k \nu + \mathcal{E}. \end{aligned} \tag{30}$$

For each $T > \nu$ consider the following set of matrices

$$\mathcal{C}(r,T) = \left\{ A \in \mathcal{C}(r)\,:\, \|A\|^2_{L_2(\Pi\otimes\mu)} \leq T \right\}$$

and the following event

$$\mathcal{B}_k = \left\{ \exists\, A \in \mathcal{C}(r,\alpha^k\nu)\,:\, \left| \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2 - \|A\|^2_{L_2(\Pi\otimes\mu)} \right| > \frac{5}{12}\alpha^k\nu + \mathcal{E} \right\}.$$

Note that $A \in S_k$ implies that $A \in \mathcal{C}(r,\alpha^k\nu)$. Then (30) implies that $\mathcal{B}_k$ holds and we obtain $\mathcal{B} \subset \cup \mathcal{B}_k$. Thus, it is enough to estimate the probability of the simpler event $\mathcal{B}_k$ and then to apply the union bound. Such an estimation is given by the following lemma. Its proof is given in Appendix C. Let

$$Z_T = \sup_{A\in\mathcal{C}(r,T)} \left| \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2 - \|A\|^2_{L_2(\Pi\otimes\mu)} \right|.$$

Lemma 5.

$$\mathbb{P}\left(Z_T > \frac{5}{12}T + \frac{44\,c_\phi^2\,l\,r}{\omega_{\min}} \left(\mathbb{E}\,\|\Sigma_R\|\right)^2 \right) \leq \exp\left(-\frac{c_3 n T^2}{c_\phi^2\, l} \right)$$

where $c_3 = \frac{1}{128}$.

Lemma 5 implies that $\mathbb{P}\left(\mathcal{B}_k\right) \leq \exp\left(-\frac{c_3\,n\,\alpha^{2k}\,\nu^2}{c_\phi^2\,l}\right)$. Using the union bound we obtain

$$\begin{aligned} \mathbb{P}\left(\mathcal{B}\right) &\leq \sum_{k=1}^{\infty} \mathbb{P}\left(\mathcal{B}_k\right) \leq \sum_{k=1}^{\infty} \exp\left(-\frac{c_3\,n\,\alpha^{2k}\,\nu^2}{c_\phi^2\,l} \right) \\ &\leq \sum_{k=1}^{\infty} \exp\left(-\frac{\left(2\,c_3\,n\log(\alpha)\,\nu^2\right)k}{c_\phi^2\,l} \right) \end{aligned}$$

where we used $e^x \geq x$. We finally compute for $\nu = c_\phi \sqrt{\frac{64\ \log(d)\, l}{\log(6/5)\, n}}$

$$\mathbb{P}\left(\mathcal{B}\right) \leq \frac{\exp\left(-\frac{2\, c_3\, n \log(\alpha)\, \nu^2}{c_\phi^2\, l}\right)}{1 - \exp\left(-\frac{2\, c_3\, n \log(\alpha)\, \nu^2}{c_\phi^2\, l}\right)} = \frac{\exp\left(-\log(d)\right)}{1 - \exp\left(-\log(d)\right)}.$$

This completes the proof of Lemma 2.

Appendix C: Proof of Lemma 5

Our approach is standard: first we show that Z_T concentrates around its expectation and then we upper bound the expectation. By definition,

$$Z_T = \sup_{A \in \mathcal{C}(r,T)} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, A\rangle^2 - \mathbb{E}\left(\langle X, A\rangle^2\right) \right|.$$

Note that

$$\left|\langle X_i, A\rangle\right| \leq \|W\|_2 \, \|\phi(t)\|_2 \, \|A\|_2 \leq c_\phi \sqrt{l},$$

where we used $\|W\|_2 \leq 1$ and condition (2).

Massart's concentration inequality (see e.g. [1, Theorem 14.2]) implies that

$$\mathbb{P}\left(Z_T \geq \mathbb{E}\left(Z_T\right) + \frac{1}{9}\frac{5}{12}T\right) \leq \exp\left(-\frac{c_3 n T^2}{c_\phi^2\, l}\right). \tag{31}$$

where $c_3 = \frac{1}{128}$.

Next we bound the expectation $\mathbb{E}\left(Z_T\right)$. Using a standard symmetrization argument (see Ledoux and Talagrand [21]) we obtain

$$\begin{aligned}\mathbb{E}\left(Z_T\right) &= \mathbb{E}\left(\sup_{A \in \mathcal{C}(r,T)} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, A\rangle^2 - \mathbb{E}\left(\langle X, A\rangle^2\right) \right|\right) \\ &\leq 2\mathbb{E}\left(\sup_{A \in \mathcal{C}(r,T)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i, A\rangle^2 \right|\right)\end{aligned}$$

where $\{\epsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence. Then, the contraction inequality (see Ledoux and Talagrand [21]) yields

$$\begin{aligned}\mathbb{E}\left(Z_T\right) &\leq 8\, c_\phi \sqrt{l}\, \mathbb{E}\left(\sup_{A \in \mathcal{C}(r,T)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i, A\rangle \right|\right) \\ &= 8\, c_\phi \sqrt{l}\, \mathbb{E}\left(\sup_{A \in \mathcal{C}(r,T)} \left|\langle \Sigma_R, A\rangle\right|\right)\end{aligned}$$

where $\Sigma_R = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_i$. For $A \in \mathcal{C}(r,T)$ we have that

$$\begin{aligned}\|A\|_* &\leq \sqrt{r}\,\|A\|_2\\ &\leq \frac{\sqrt{r}\,\|A\|_{L_2(\Pi\otimes\mu)}}{\sqrt{\omega_{\min}}}\\ &\leq \sqrt{\frac{r\,T}{\omega_{\min}}}\end{aligned}$$

where we have used (10).

Then, by duality between nuclear and operator norms, we compute

$$\begin{aligned}\mathbb{E}\left(Z_T\right) &\leq 8\,c_\phi\,\sqrt{l}\,\mathbb{E}\left(\sup_{\|A\|_*\leq\sqrt{r\,T/\omega_{\min}}}|\langle\Sigma_R, A\rangle|\right)\\ &\leq 8\,c_\phi\,\sqrt{\frac{l\,r\,T}{\omega_{\min}}}\,\mathbb{E}\left(\|\Sigma_R\|\right).\end{aligned}$$

Finally, using

$$\frac{1}{9}\frac{5}{12}T + 8\,c_\phi\sqrt{\frac{l\,r\,T}{\omega_{\min}}}\,\mathbb{E}\left(\|\Sigma_R\|\right) \leq \left(\frac{1}{9}+\frac{8}{9}\right)\frac{5}{12}T + \frac{44\,c_\phi^2\,l\,r}{\omega_{\min}}\left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2$$

and the concentration bound (31) we obtain that

$$\mathbb{P}\left(Z_T > \frac{5}{12}T + \frac{44\,c_\phi^2\,l\,r}{\omega_{\min}}\left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2\right) \leq \exp\left(-\frac{c_3 n T^2}{c_\phi^2\,l}\right)$$

where $c_3 = \frac{1}{128}$ as stated.

Appendix D: Proof of Lemma 3

Using (19) we compute

$$\lambda\left(\|\hat{A}\|_1 - \|A_0\|_1\right) \leq 2\,\|\Sigma\|\,\|H\|_1.$$

The condition $\lambda \geq 3\,\|\Sigma\|$, the triangle inequality and (21) yield

$$\lambda\left(\left\|\mathbf{P}_{A_0}^{\perp}(H)\right\|_1 - \|\mathbf{P}_{A_0}(H)\|_1\right) \leq \frac{2}{3}\lambda\left(\left\|\mathbf{P}_{A_0}^{\perp}(H)\right\|_1 + \|\mathbf{P}_{A_0}(H)\|_1\right).$$

This implies that

$$\left\|\mathbf{P}_{A_0}^{\perp}(H)\right\|_1 \leq 5\,\|\mathbf{P}_{A_0}(H)\|_1\,.$$

as stated.

Appendix E: Bounds on the stochastic errors

In this section we will obtain upper bounds for the stochastic errors $\|\Sigma\|$, $\|\Sigma_R\|$. Recall that

$$\Sigma_R = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_i \qquad \text{and} \qquad \Sigma = \frac{1}{n}\sum_{i=1}^{n}\left(W_i^T \rho^{(l)}(t_i) + \sigma\, \xi_i\right) X_i \tag{32}$$

where $\{\epsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence.

The following proposition is the matrix version of Bernstein's inequality in the bounded case (see Theorem 1.6 in [25]). Let $Z_1, \dots, Z_n$ be independent random matrices with dimensions $m_1 \times m_2$. Define

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

Proposition 1. *Let $Z_1, \dots, Z_n$ be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}(Z_i) = 0$. Suppose that $\|Z_i\| \le U$ for some constant U and all $i = 1, \dots, n$. Then, for all $t > 0$, with probability at least $1 - e^{-t}$ we have*

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\| \le 2\max\left\{\sigma_Z\sqrt{\frac{t+\log(d)}{n}}, U\,\frac{t+\log(d)}{n}\right\},$$

where $d = m_1 + m_2$.

It is possible to extend this result to the sub-exponential case. Set

$$U_i = \inf\left\{K > 0 \;:\; \mathbb{E}\exp\left(\|Z_i\|/K\right) \le e\right\}.$$

The following proposition is obtained by an extension of Theorem 4 in [18] to rectangular matrices via self-adjoint dilation (cf., for example 2.6 in [25]).

Proposition 2. *Let $Z_1, \dots, Z_n$ be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}(Z_i) = 0$. Suppose that $U_i < U$ for some constant U and all $i = 1, \dots, n$. Then, there exists an absolute constant c^*, such that, for all $t > 0$, with probability at least $1 - e^{-t}$ we have*

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\| \le c^*\max\left\{\sigma_Z\sqrt{\frac{t+\log(d)}{n}}, U\left(\log\frac{U}{\sigma_Z}\right)\frac{t+\log(d)}{n}\right\},$$

where $d = m_1 + m_2$.

We use Propositions 1 and 2 to prove Lemmas 1 and 4.

E.1. Proof of Lemma 1

Let $\Sigma_1 = \frac{1}{n}\sum\limits_{i=1}^{n} W_i^T \rho^{(l)}(t_i) X_i$ and $\Sigma_2 = \frac{1}{n}\sum\limits_{i=1}^{n}\xi_i X_i$. Then, we obtain $\Sigma = \Sigma_1 + \sigma\,\Sigma_2$. In order to derive an upper bound for $\|\Sigma_2\|$, we apply Proposition 2 to

$$Z_i = \xi_i X_i = \xi_i W_i \phi^T(t_i).$$

We need to estimate σ_Z and U. Note that Z_i is a zero-mean random matrix such that

$$\begin{aligned}\|Z_i\| &\leq |\xi_i| \left\|W_i\phi^T(t_i)\right\|_2 = |\xi_i| \left\|W_i\phi^T(t_i)\right\|_2\\ &= |\xi_i| \left\|W_i\right\|_2 \left\|\phi^T(t_i)\right\|_2 \leq |\xi_i| \left\|\phi^T(t_i)\right\|_2\\ &\leq |\xi_i| \, c_\phi \sqrt{l}\end{aligned}$$

where we used condition (2) and $\|W\|_2 \leq 1$. Then, Assumption 4 implies that there exists a constant K such that $U_i \leq K\, c_\phi \sqrt{l}$ for all $i = 1, \dots, n$.

Let us estimate σ_Z for $Z = \xi\, W\phi^T(t)$. First we compute $\frac{1}{n}\sum\limits_{i=1}^{n}\mathbb{E}\left(Z_i\, Z_i^T\right)$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_i Z_i^T\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\xi_i^2 W_i\phi^T(t_i)\phi(t_i)W_i^T\right)\\ &= \mathbb{E}\left(\|\phi(t)\|_2^2\, W\, W^T\right)\\ &= l\,\Omega\end{aligned} \tag{33}$$

where we used $\mathbb{E}(\xi^2) = 1$.

Now we compute $\frac{1}{n}\sum\limits_{i=1}^{n}\mathbb{E}\left(Z_i^T\, Z_i\right)$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_i^T Z_i\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\xi_i^2 \phi(t_i)W_i^T W_i\phi^T(t_i)\right)\\ &= \mathbb{E}\left(\phi(t)\phi^T(t)\,\|W\|_2^2\right)\\ &= \mathrm{tr}\left(\Omega\right)\,\mathbb{I}_l\end{aligned} \tag{34}$$

where we used that $(\phi_i(\cdot))_{i=1,\dots,\infty}$ is an orthonormal basis in $L_2\left((0,1), d\mu\right)$.

Equations (33) and (34) imply that

$$\sigma_Z^2 \leq (l\,\omega_{\max}) \vee \mathrm{tr}\left(\Omega\right) \quad \text{and} \quad \sigma_Z^2 \geq l\,\omega_{\max}.$$

Applying Proposition 2 we derive that for all $t > 0$ with probability at least $1 - e^{-t}$

$$\|\Sigma_2\| \leq c^* \max\left\{\sqrt{\frac{M\ (t + \log(d))}{n}}, \frac{K\, c_\phi \sqrt{l}\ (t + \log(d))\ \log\left(\frac{K\, c_\phi}{\omega_{\max}}\right)}{n}\right\} \tag{35}$$

where $M = \mathrm{tr}(\Omega) \vee (l\omega_{\max})$.

One can estimate $\|\Sigma_1\|$ in a similar way. We apply Proposition 1 to

$$\begin{aligned}Z_i &= W_i^T\rho^{(l)}(t_i)X_i\\ &= W_i^T\rho^{(l)}(t_i)W_i\phi^T(t_i).\end{aligned}$$

We begin by proving that

$$\mathbb{E}\left(W^T\rho^{(l)}(t)W\phi^T(t)\right)=0.$$

Let $W=(w_1,\dots,w_p)$. The (m,k)-th entry of $W^T\rho^{(l)}(t)W\phi^T(t)$ is equal to $\sum\limits_{j=1}^{p} w_j\,\rho_j^{(l)}(t)\,w_m\,\phi_k(t)$. By definition $\rho_j^{(l)}(t)=f_j(t)-\sum\limits_{i=1}^{l}a_{ji}^0\phi_i(t)$ and we compute

$$\begin{aligned}\mathbb{E}\left(\rho_j^{(l)}(t)\phi_k(t)\right)&=\mathbb{E}\left(\left(f_j(t)-\sum_{i=1}^{l}a_{ji}^0\phi_i(t)\right)\phi_k(t)\right)\\&=\mathbb{E}\left(f_j(t)\phi_k(t)-\sum_{i=1}^{l}a_{ji}^0\phi_i(t)\phi_k(t)\right)\\&=a_{jk}^0-a_{jk}^0=0\end{aligned}$$

since $(\phi_i(\cdot))_{i=1,\dots,\infty}$ is an orthonormal basis. Therefore,

$$\begin{aligned}\mathbb{E}\left(\sum_{j=1}^{p}w_j\,\rho_j^{(l)}(t)\,w_m\,\phi_k(t)\right)&=\sum_{j=1}^{p}\mathbb{E}_W\left(w_j\,w_m\mathbb{E}_t\left(\rho_j^{(l)}(t)\,\phi_k(t)\right)\right)\\&=0\end{aligned}$$

Next we estimate U. Note that $\rho^{(l)}(t)$ has at most $s-1$ non-zero coefficients. Then, Assumption 1 and $\|W\|_2\le 1$ imply that $t-$almost surely $\left(W^T\rho^{(l)}(t)\right)^2\le\frac{b^2\,(s-1)}{l^{2\gamma}}$ and

$$\begin{aligned}\|Z_i\|&\le|W_i^T\rho^{(l)}(t_i)|\,\left\|W_i\phi^T(t_i)\right\|\\&\le\frac{b\,c_\phi\,\sqrt{l\,(s-1)}}{l^\gamma}.\end{aligned}$$

Let us estimate σ_Z for $Z=\left(W^T\rho^{(l)}(t)\right)W\phi^T(t)$. First we compute $\frac{1}{n}\sum\limits_{i=1}^{n}\mathbb{E}\left(Z_i\,Z_i^T\right)$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_iZ_i^T\right)&=\mathbb{E}\left(\left(W^T\rho^{(l)}(t)\right)^2W\phi^T(t)\phi(t)W^T\right)\\&=\mathbb{E}_t\left(\|\phi(t)\|_2^2\;\mathbb{E}_W\left(\left(W^T\rho^{(l)}(t)\right)^2WW^T\right)\right).\end{aligned}$$

We obtain

$$\mathbb{E}_W\left(\left(W^T\rho^{(l)}(t)\right)^2WW^T\right)\le\frac{b^2\,(s-1)}{l^{2\gamma}}\,\mathbb{E}\left(WW^T\right)$$

where we used $WW^T\ge 0$. Finally we obtain

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_iZ_i^T\right)\right\|\le\frac{b^2\,(s-1)\,\omega_{\max}\,l}{l^{2\gamma}}. \tag{36}$$

Now we compute $\frac{1}{n}\sum\limits_{i=1}^{n}\mathbb{E}\left(Z_i^T\, Z_i\right)$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_i^T Z_i\right) = \mathbb{E}_t\left(\left(W^T\rho^{(l)}(t)\right)^2 \phi(t)W^T W\phi^T(t)\right)$$
$$= \mathbb{E}_t\left(\left(W^T\rho^{(l)}(t)\right)^2 \|W\|_2^2\,\phi(t)\phi^T(t)\right).$$

Using $\mathbb{E}(\|W\|_2^2) = \mathrm{tr}(\Omega)$ and $\mathbb{E}_t\left(\phi(t)\phi^T(t)\right) = \mathbb{I}_l$ we obtain

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_i^T Z_i\right)\right\| \leq \frac{b^2\,(s-1)}{l^{2\gamma}}\,\mathrm{tr}(\Omega). \tag{37}$$

Equations (36) and (37) imply that

$$\sigma_Z^2 \leq \frac{b^2\,(s-1)}{l^{2\gamma}}\left[\mathrm{tr}(\Omega)\vee(l\,\omega_{\max})\right].$$

Applying Proposition 1, we derive that for all $t>0$ with probability at least $1-e^{-t}$

$$\|\Sigma_1\| \leq \frac{2\,b\,\sqrt{s-1}}{l^{\gamma}}\max\left\{\sqrt{\frac{M(t+\log(d))}{n}},\,\frac{c_\phi\,\sqrt{l}\,(t+\log(d))}{n}\right\}. \tag{38}$$

The bounds (38) and (35) imply that for all $t>0$ with probability at least $1-2e^{-t}$

$$\|\Sigma\| \leq \left(\sigma\,c^* + \frac{2\,b\,\sqrt{s-1}}{l^{\gamma}}\right)\max\left\{\sqrt{\frac{M\,(t+\log(d))}{n}},\right.$$
$$\left.\frac{c_\phi\,\sqrt{l}\;(t+\log(d))\left(\left[K\,\log\left(\frac{K}{\omega_{\max}}\right)\right]\vee 1\right)}{n}\right\}$$

as stated.

E.2. Proof of Lemma 4

The proof follows the lines of the proof of Lemma 7 in [17]. We use Proposition 1 with $Z_i = \epsilon_i\,X_i$. As in the proof of Lemma 1, we obtain $U=\sqrt{l}$ and $\sigma_Z^2 = (\mathrm{tr}(\Omega)\vee(l\sigma_{\max}(\Omega)))$. Set $M = (\mathrm{tr}(\Omega)\vee(l\sigma_{\max}(\Omega)))$, then Proposition 1 implies that for all $t>0$ with probability at least $1-e^{-t}$

$$\|\Sigma_R\| \leq 2\max\left\{\sqrt{\frac{M\,(t+\log(d))}{n}},\,\frac{\sqrt{l}\,(t+\log(d))}{n}\right\}. \tag{39}$$

Set $t^* = \frac{n\,M}{l} - \log(d)$ so that t^* is the value of t such that the two terms in (39) are equal. Note that (39) implies that

$$\mathbb{P}\left(\|\Sigma_R\| > t\right) \leq d \exp\left\{-\frac{t^2\,n}{4\,M}\right\} \qquad \text{for} \qquad t \leq t^* \tag{40}$$

and

$$\mathbb{P}\left(\|\Sigma_R\| > t\right) \leq d \exp\left\{-\frac{t\,n}{2\,\sqrt{l}}\right\} \qquad \text{for} \qquad t \geq t^*. \tag{41}$$

We set $\nu_1 = \frac{n}{4\,M}$, $\nu_2 = \frac{n}{2\sqrt{l}}$. By Hölder's inequality we derive

$$\mathbb{E}\,\|\Sigma_R\| \leq \left(\mathbb{E}\,\|\Sigma_R\|^{2\log(d)}\right)^{1/(2\log(d))}.$$

Inequalities (40) and (41) imply that

$$\begin{aligned}\left(\mathbb{E}\,\|\Sigma_R\|^{2\log(d)}\right)^{1/2\log(d)} &= \left(\int\limits_0^{+\infty} \mathbb{P}\left(\|\Sigma_R\| > t^{1/(2\log(d))}\right) \mathrm{d}t\right)^{1/2\log(d)} \\ &\leq \left(d\int\limits_0^{+\infty} \exp\{-t^{1/\log(d)}\nu_1\}\mathrm{d}t + d\int\limits_0^{+\infty} \exp\{-t^{1/(2\log(d)}\nu_2\}\mathrm{d}t\right)^{1/2\log(d)} \\ &\leq \sqrt{e}\left(\log(d)\nu_1^{-\log(d)}\Gamma(\log(d)) + 2\log(d)\;\nu_2^{-2\log(d)}\Gamma(2\log(d))\right)^{1/(2\log(d))}.\end{aligned} \tag{42}$$

Recall that Gamma-function satisfies the following inequality

$$\Gamma(x) \leq \left(\frac{x}{2}\right)^{x-1} \quad \text{for} \quad x \geq 2, \tag{43}$$

(see e.g. [17]). Plugging (43) into (42) we compute

$$\begin{aligned}\mathbb{E}\,\|\Sigma_R\| \leq \sqrt{e}\,\Big(&(\log(d))^{\log(d)}\nu_1^{-\log(d)}2^{1-\log(d)} \\ &+ 2(\log(d))^{2\log(d)}\nu_2^{-2\log(d)}\Big)^{1/(2\log(d))}.\end{aligned}$$

Observe that $n \geq n^*$ implies $\nu_1 \log(d) \leq \nu_2^2$ and we obtain

$$\mathbb{E}\,\|\Sigma_R\| \leq \sqrt{\frac{2e\log(d)}{\nu_1}}. \tag{44}$$

We conclude the proof by plugging $\nu_1 = \frac{n}{4\,M}$ into (44).

References

[1] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer. MR2807761

[2] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, **9(6)**, 717–772. MR2565240

[3] 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

[4] 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.

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

[6] Fan, J., and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1**, 179–195. MR2425354

[7] 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. MR1229881

[8] 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

[9] 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

[10] 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. MR2101537

[11] 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. MR2087972

[12] Kauermann, G. and Tutz, G. (1999). On model diagnostics using varying coefficient models. *Biometrika*, **86**, 119–128. MR1688076

[13] 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. MR2797848

[14] Keshavan, R. H., Montanari, A. and Oh, S. (2010). Matrix completion from a few entries. *IEEE Trans. on Info. Th.*, **56(6)**, 2980–2998. MR2683452

[15] Klopp, O. (2011). Matrix completion with unknown variance of the noise. `http://arxiv.org/abs/1112.3055`

[16] Klopp, O. (2012). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, to appear.

[17] Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, **5**, 1161–1183. MR2842903

[18] Koltchinskii, V. (2011). A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. *IMS Collections, Festschritt in Honor of J. Wellner.*

[19] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, **39(5)**, 2302–2329. MR2906869

[20] Lian, H. (2012). Spline Estimator for Simultaneous Variable Selection and Constant Coefficient Identification in High-dimensional Generalized Varying-Coefficient Models. Manuscript.

[21] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, New York, NY. MR1102015

[22] Mallat, S. (2009). *A Wavelet Tour of Signal Processing,* Third Ed., Elsevier, New York. MR2479996

[23] Negahban, S. and Wainwright, M. J. (2010). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, **13**, 1665–1697. MR2930649

[24] Senturk, D. and Mueller, H. G. (2010). Functional varying coefficient models for longitudinal data. *J. Amer. Statist. Assoc.*, **105**, 1256–1264. MR2752619

[25] Tropp, J. A. (2011). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, **11(4)**. MR2946459

[26] Tsybakov, A. (2010). *Introduction to Nonparametric Estimation*, Springer Series in Statistics. MR2013911

[27] Wang, L., Kai, B., and Li, R. (2009). Local Rank Inference for Varying Coefficient Models. *J. Amer. Statist. Assoc.*, **104**, 1631–1645. MR2597005

[28] Wu, C. O., Chiang, C. T. and Hoover, D. R. (1998). Asymptotic confi- dence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.*, **93**, 1388–1402. MR1666635

[29] 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. MR2328308