ANZIAM J. 58(2017), 220–230
doi:10.1017/S1446181116000328

LEARNING GRADIENTS FROM NONIDENTICAL DATA

XUE-MEI DONG[1]

(Received 11 August, 2016; accepted 17 September, 2016; first published online 7 March 2017)

Abstract

Selecting important variables and estimating coordinate covariation have received considerable attention in the current big data deluge. Previous work shows that the gradient of the regression function, the objective function in regression and classification problems, can provide both types of information. In this paper, an algorithm to learn this gradient function is proposed for nonidentical data. Under some mild assumptions on data distribution and the model parameters, a result on its learning rate is established which provides a theoretical guarantee for using this method in dynamical gene selection and in network security for recognition of malicious online attacks.

2010 *Mathematics subject classification*: primary 68T05; secondary 62J02.

Keywords and phrases: gradient learning, variable selection, coordinate covariation estimation.

1. Introduction

With the increasing collection of vast quantities of data, it has become common to encounter high-dimensional data sets in a variety of applications. In general, a complicated model including many insignificant variables may result in less predictive ability. Hence, it is desirable to select some important variables and estimate coordinate covariance [7, 13].

Variable selection, also known as feature screening, aims at choosing a subset of variables most relevant for predicting responses. Using a variety of criteria, for example correlation or information theory, to rank features is a common way. The variables with scores below a threshold are eliminated [4]. These ranking-based methods focus on individual prediction power and are ineffective in selecting a subset of variables that are marginally weak but in combination strong in prediction. Bayesian learning [8] is another popular method based on some prior information, such as sparsity. Lasso [12] and elastic net [14] are two widely used approaches of this type.

[1] School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, 310018, China; e-mail: dongxuemei@zjgsu.edu.cn.

However, they are based on the assumption of a linear model which is not suitable for many practical applications.

Coordinate covariance is a measure of how much two random variables change together. A sample covariance matrix is not a good estimator of population covariance if the dimension of the input variable is high. Many methods have been proposed to estimate the covariance in this case [2, 3]. However, almost all of them made some assumptions on the distribution, such as a Gaussian distribution for simplicity or an exponential distribution for sparsity. Moreover, these approaches made no connection with variable selection.

Mukherjee and Zhou [7] introduced a gradient-learning algorithm which can provide both variable selection and a covariance estimate at the same time. The motivation of this algorithm is that the gradient of the prediction function provides a natural interpretation of the geometric structure of the data. The larger the norm of the partial derivative with respect to a variable, the more important the corresponding variable is likely to be for prediction. Also the inner product between partial derivatives indicates the coordinate covariance with respect to variation in the prediction function. However, the data in that work needed to be sampled from an unknown independent and identical distribution (i.i.d.). In many application domains, this i.i.d. condition becomes inappropriate. For example, there are many Markov models which are not identical [5, 11], and some discrete contracting dynamical systems are not i.i.d. [9]. In this work, Under a Markov sampling condition [10], we establish an online gradient-learning algorithm. The method is based on a nonlinear model, and there are no special distribution assumptions for the samples.

The rest of the paper is organized as follows. In Section 2, the definition of the proposed algorithm is given and some supporting results are introduced. Error analysis of this algorithm is presented in Section 3, which ensures the feasibility of using this algorithm theoretically.

2. Notation and definitions

We first introduce some notation that will be used in Section 3, and give the definition of our proposed algorithm.

Let the input space X be a compact subset of $\mathbb{R}^n$ and the output space be $Y = [-M, M]$ for some $M > 0$. Each $\mathbf{x} \in X$ is assigned a conditional probability measure $\rho(\cdot|\mathbf{x})$ on Y. The regression function is defined as

$$f_\rho(\mathbf{x}) = \int_Y y \, d\rho(y|\mathbf{x}) \quad \mathbf{x} \in X.$$

Define $\mathbf{x} = (x^1, \dots, x^n) \in \mathbb{R}^n$ and denote the gradient of the regression function by $\nabla f_\rho = (\partial f_\rho/\partial x^1, \partial f_\rho/\partial x^2, \dots, \partial f_\rho/\partial x^n)^T$. Our goal is to obtain ∇f_ρ from data $\{\mathbf{z}_t\}_{t\geq 1}$ with each $\mathbf{z}_t = (\mathbf{x}_t, y_t)$ independently drawn from a probability distribution $\rho^{(t)}$ defined on the product space $Z = X \times Y$.

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

Under the identical distribution assumption, namely, $\rho^{(t)}$ being fixed as ρ for each t, Mukherjee and Zhou [7] proposed a least-squares type learning algorithm, defined by

$$\mathbf{f}_{S,\lambda} = \operatorname*{argmin}_{\mathbf{f}\in\mathcal{H}_K^n}\bigg[\frac{1}{T^2}\sum_{i,j=1}^{T} w_{i,j}^{(\sigma)}\{y_i - y_j + \mathbf{f}(\mathbf{x}_i)\cdot(\mathbf{x}_j - \mathbf{x}_i)\}^2 + \lambda\|\mathbf{f}\|^2_{\mathcal{H}_K^n}\bigg], \tag{2.1}$$

where λ, σ are two positive constants called the *regularization parameters*. Let $S = \{\mathbf{z}_i\}_{i=1}^T$ and let "$\cdot$" denote the inner product in $\mathbb{R}^n$. The weight $w_{i,j}^{\sigma} = w_{\mathbf{x}_i,\mathbf{x}_j}^{\sigma} = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\}$ is used to govern the "nearness" of the samples $\mathbf{x}_i$ and $\mathbf{x}_j$. The function $K : X \times X \to \mathbb{R}$ is a Mercer kernel [1] and $\mathcal{H}_K$ is the *reproducing kernel Hilbert space* (RKHS) associated with K. The hypothesis space $\mathcal{H}_K^n$ in equation (2.1) is the n-fold of $\mathcal{H}_K$, consisting of vectors of functions $\mathbf{f} = (f^1, f^2, \ldots, f^n)^T$ with the norm $\|\mathbf{f}\|_{\mathcal{H}_K^n} = \{\sum_{\ell=1}^n \|f^\ell\|_K^2\}^{1/2}$. Denote $\kappa = \sup_{\mathbf{x},\mathbf{r}\in X}\sqrt{K(\mathbf{x},\mathbf{r})}$.

The purpose of this paper is to establish an algorithm for gradient learning with the identical distribution assumption used in (2.1) weakened. As illustrated by Smale and Zhou [10], we suppose that the marginal distribution sequence $\{\rho_X^{(t)}\}$ of $\{\rho^{(t)}\}$ converges exponentially fast in the dual of the Hölder space $C^s(X)$ for $0 < s \le 1$ [6]. Here the Hölder space $C^s(X)$ is defined as the space of all continuous functions on X with the norm $\|f\|_{C^s(X)} = \|f\|_{C(X)} + |f|_{C^s(X)}$ being finite, where $|f|_{C^s(X)} = \sup_{\mathbf{x}\neq\mathbf{r}\in X}(|f(\mathbf{x}) - f(\mathbf{r})|/\|\mathbf{x} - \mathbf{r}\|^s)$.

Definition 2.1. For $0 < s \le 1$, we say that the sequence $\{\rho_X^{(t)}\}$ converges to a probability measure ρ_X exponentially fast in $(C^s(X))^*$ if there exist $C > 0$, $0 < \alpha < 1$ and a probability measure ρ_X defined on X such that, for any $f \in C^s(X)$,

$$\bigg|\int_X f(\mathbf{x})\,d\rho_X^{(t)}(\mathbf{x}) - \int_X f(\mathbf{x})\,d\rho_X(\mathbf{x})\bigg| \le C\alpha^t\|f\|_{C^s(X)} \quad \text{for all } t\in\mathbb{N}. \tag{2.2}$$

Now we define our gradient-learning algorithm as follows.

Definition 2.2. The online algorithm for learning the gradient of the regression function with $\mathbf{z}_t$ sampling from $\rho^{(t)}$ independently is defined by $\mathbf{f}_1 = 0$ and, for $t \in \mathbb{N}$,

$$\mathbf{f}_{t+1} = \mathbf{f}_t - \eta_t[w_{2t-1,2t}^{\sigma}\{y_{2t-1} - y_{2t} + \mathbf{f}_t(\mathbf{x}_{2t-1})\cdot(\mathbf{x}_{2t} - \mathbf{x}_{2t-1})\}(\mathbf{x}_{2t} - \mathbf{x}_{2t-1})K_{\mathbf{x}_{2t-1}} + \lambda_t\mathbf{f}_t]. \tag{2.3}$$

Here η_t and λ_t denote the step sizes and regularization parameters, respectively.

For brevity, we define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K^n \longrightarrow \mathbb{R}^n$ as $S_{\mathbf{x}}(\mathbf{f}) = \mathbf{f}(\mathbf{x}) = (f^1(\mathbf{x}), \ldots, f^n(\mathbf{x}))^T$ and its adjoint operator as $S_{\mathbf{x}}^T : \mathbb{R}^n \longrightarrow \mathcal{H}_K^n$, with $S_{\mathbf{x}}^T(\mathbf{c}) = \mathbf{c}K_{\mathbf{x}}$, for all $\mathbf{c} \in \mathbb{R}^n$. Denote

$$Y_{\mathbf{x}_{2t}} = w_{2t-1,2t}^{\sigma}(y_{2t} - y_{2t-1})(\mathbf{x}_{2t} - \mathbf{x}_{2t-1}) \in \mathbb{R}^n,$$
$$D_{\mathbf{x}_{2t}} = w_{2t-1,2t}^{\sigma}(\mathbf{x}_{2t} - \mathbf{x}_{2t-1})(\mathbf{x}_{2t} - \mathbf{x}_{2t-1})^T \in \mathbb{R}^{n\times n}.$$

We can rewrite our online algorithm (2.3) as

$$\mathbf{f}_{t+1} = (1 - \eta_t\lambda_t)\mathbf{f}_t - \eta_t\{S_{\mathbf{x}_{2t-1}}^T D_{\mathbf{x}_{2t}} S_{\mathbf{x}_{2t-1}}(\mathbf{f}_t) - S_{\mathbf{x}_{2t-1}}^T Y_{\mathbf{x}_{2t}}\}.$$

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

To analyse the learning rate of $\mathbf{f}_{t+1}$ to ∇f_ρ, we introduce the integral operators

$$L_{K,\rho^{(2t)}}(\mathbf{f}) = \int_X \int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^T K_{\mathbf{x}}\mathbf{f}(\mathbf{x})\, d\rho_X^{(2t-1)}(\mathbf{x})\, d\rho_X^{(2t)}(\mathbf{u}),$$
$$L_{K,\rho}(\mathbf{f}) = \int_X \int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^T K_{\mathbf{x}}\mathbf{f}(\mathbf{x})\, d\rho_X(\mathbf{x})\, d\rho_X(\mathbf{u}).$$

Define the vector functions

$$\mathbf{f}_{\lambda,\rho^{(2t)}} = (L_{K,\rho^{(2t)}} + \lambda I)^{-1}\mathbf{f}_{\rho^{(2t)},\sigma}, \tag{2.4}$$
$$\mathbf{f}_{\lambda,\rho} = (L_{K,\rho} + \lambda I)^{-1}\mathbf{f}_{\rho,\sigma}, \tag{2.5}$$

where

$$\mathbf{f}_{\rho^{(2t)},\sigma} = \int_X \int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(f_\rho(\mathbf{u}) - f_\rho(\mathbf{x}))K_{\mathbf{x}}\, d\rho_X^{(2t-1)}(\mathbf{x})\, d\rho_X^{(2t)}(\mathbf{u}),$$
$$\mathbf{f}_{\rho,\sigma} = \int_X \int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(f_\rho(\mathbf{u}) - f_\rho(\mathbf{x}))K_{\mathbf{x}}\, d\rho_X(\mathbf{x})\, d\rho_X(\mathbf{u}).$$

The error between $\mathbf{f}_{t+1}$ and ∇f_ρ can be decomposed into three parts as

$$\mathbf{f}_{t+1} - \nabla f_\rho = \{\mathbf{f}_{t+1} - \mathbf{f}_{\lambda_t,\rho^{(2t)}}\} + \{\mathbf{f}_{\lambda_t,\rho^{(2t)}} - \mathbf{f}_{\lambda_t,\rho}\} + \{\mathbf{f}_{\lambda_t,\rho} - \nabla f_\rho\}. \tag{2.6}$$

The first part of right-hand side is referred to as *sample error*, and the third part as *approximation error*. The second part is caused by different measures. These errors will be estimated in the next section.

For convergence analysis, the conditional distribution $\{\rho_X^{(t)}(y|\mathbf{x})\}$ of $\{\rho_X^{(t)}\}$ is assumed to be independent of t and is denoted by $\rho_{\mathbf{x}}$. Furthermore, we need the kernel K to satisfy the *kernel condition*, that is, $K \in C^s(X \times X)$ and there exists $\kappa_s > 0$ such that, for all $\mathbf{x}, \mathbf{u} \in X$,

$$\|K_{\mathbf{x}} - K_{\mathbf{u}}\|_K \le \kappa_s(\|\mathbf{x}-\mathbf{u}\|^s),$$

where $K_{\mathbf{x}} = K(\mathbf{x}, \cdot)$. The Mercer kernel K with the kernel condition yields that $\|f\|_{C^s(X)} \le (\kappa + \kappa_s)\|f\|_K$ for any $f \in \mathcal{H}_K$. Therefore $f \in C^s(X)$.

3. Main results

The first contribution of this paper is to estimate the second part in right-hand side of (2.6), as follows.

THEOREM 3.1. *Let $\mathbf{f}_{\lambda,\rho^{(2t)}}$ and $\mathbf{f}_{\lambda,\rho}$ be given by (2.4) and (2.5), respectively. Assume that, for the conditional distributions $\{\rho_{\mathbf{x}} : \mathbf{x} \in X\}$, there exists a constant $C_\rho \ge 0$ such that $\|\rho_{\mathbf{x}} - \rho_{\mathbf{u}}\|_{(C^s(Y))^*} \le C_\rho\|\mathbf{x}-\mathbf{u}\|^s$ for all $\mathbf{x}, \mathbf{u} \in X$. Then,*

$$\|\mathbf{f}_{\lambda,\rho} - \mathbf{f}_{\lambda,\rho^{(2t)}}\|_{\mathcal{H}_K^n} \le \widetilde{C}_0\alpha^{2t-1}\lambda^{-3/2}\sigma^{-1},$$

where the constant $\widetilde{C}_0$ is independent of α, σ and λ.

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

Proof. Using (2.4) and (2.5),

$$\begin{aligned}\mathbf{f}_{\lambda,\rho}-\mathbf{f}_{\lambda,\rho^{(2t)}}&=(L_{K,\rho}+\lambda I)^{-1}\{\mathbf{f}_{\rho,\sigma}-(L_{K,\rho}+\lambda I)\mathbf{f}_{\lambda,\rho^{(2t)}}\}\\&=(L_{K,\rho}+\lambda I)^{-1}\{(\mathbf{f}_{\rho,\sigma}-\mathbf{f}_{\rho^{(2t)},\sigma})+(L_{K,\rho^{(2t)}}-L_{K,\rho})\mathbf{f}_{\lambda,\rho^{(2t)}}\}.\end{aligned}$$

First, we rewrite this as

$$\begin{aligned}\mathbf{f}_{\rho,\sigma}-\mathbf{f}_{\rho^{(2t)},\sigma}&=\int_X\int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(f_\rho(\mathbf{u})-f_\rho(\mathbf{x}))K_x\,d\rho_X(\mathbf{x})\,d(\rho_X-\rho_X^{(2t)})(\mathbf{u})\\&\quad+\int_X\int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(f_\rho(\mathbf{u})-f_\rho(\mathbf{x}))K_x\,d\rho_X^{(2t)}(\mathbf{u})\,d(\rho_X-\rho_X^{(2t-1)})(\mathbf{x})\\&=\mathrm{I}_a+\mathrm{I}_b.\end{aligned}$$

Let $\mathbf{h}(\mathbf{x})=\int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(f_\rho(\mathbf{u})-f_\rho(\mathbf{x}))\,d(\rho_X-\rho_X^{(2t)})(\mathbf{u})$ and let $h^i(\mathbf{x})$ be its ith coordinate. By the reproducing property

$$\|\mathrm{I}_a\|^2_{\mathcal{H}_K^n}=\sum_{i=1}^n\bigg\|\int_X h^i(\mathbf{x})K_{\mathbf{x}}\,d\rho_X(\mathbf{x})\bigg\|_K^2=\sum_{i=1}^n\int_X\int_X h^i(\mathbf{x})K(\mathbf{x},\mathbf{r})h^i(\mathbf{r})\,d\rho_X(\mathbf{x})\,d\rho_X(\mathbf{r}).$$

According to Fubini's theorem,

$$\begin{aligned}\|\mathrm{I}_a\|^2_{\mathcal{H}_K^n}=\sum_{i=1}^n\int_X\bigg[\int_X\int_X& h^i(\mathbf{x})K(\mathbf{x},\mathbf{r})w_{\mathbf{r},\boldsymbol{\tau}}(\boldsymbol{\tau}^i-\mathbf{r}^i)(f_\rho(\boldsymbol{\tau})\\&-f_\rho(\mathbf{r}))\,d\rho_X(\mathbf{x})\,d\rho_X(\mathbf{r})\bigg]\,d(\rho_X-\rho_X^{(2t)})(\boldsymbol{\tau}),\end{aligned}$$

which, together with (2.2), gives

$$\begin{aligned}\|\mathrm{I}_a\|^2_{\mathcal{H}_K^n}\le nC\alpha^{2t}\max_{1\le i\le n}\bigg\|\int_X\int_X& h^i(\mathbf{x})K(\mathbf{x},\mathbf{r})w_{\mathbf{r},\boldsymbol{\tau}}(\boldsymbol{\tau}^i-\mathbf{r}^i)(f_\rho(\boldsymbol{\tau})\\&-f_\rho(\mathbf{r}))\,d\rho_X(\mathbf{x})\,d\rho_X(\mathbf{r})\bigg\|_{C^s(X)}.\end{aligned}$$

Let $\|\cdot\|_{\mathbf{u},C^s(X)}$ and $\|\cdot\|_{\mathbf{u},C(X)}$ be the norms with respect to $\mathbf{u}$, and $D=\text{diameter}(X)$. Then

$$\begin{aligned}\|w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}^i-\mathbf{x}^i)(f_\rho(\mathbf{u})-f_\rho(\mathbf{x}))\|_{\mathbf{u},C^s(X)}&\le\|w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}^i-\mathbf{x}^i)(f_\rho(\mathbf{u})-f_\rho(\mathbf{x}))\|_{\mathbf{u},C(X)}\\&\quad+\|w^{\sigma}_{\mathbf{x},\mathbf{u}}(f_\rho(\mathbf{u})-f_\rho(\mathbf{x}))\|_{\mathbf{u},C(X)}|\mathbf{u}^i-\mathbf{x}^i|_{\mathbf{u},C^s(X)}\\&\quad+|w^{\sigma}_{\mathbf{x},\mathbf{u}}(f_\rho(\mathbf{u})-f_\rho(\mathbf{x}))|_{\mathbf{u},C^s(X)}\|\mathbf{u}^i-\mathbf{x}^i\|_{\mathbf{u},C(X)}\\&\le 2MD+2MD^{1-s}+D(2M|w^{\sigma}_{\mathbf{x},\mathbf{u}}|_{\mathbf{u},C^s(X)}\\&\quad+|f_\rho(\mathbf{u})-f_\rho(\mathbf{x})|_{\mathbf{u},C^s(X)})\\&\le 2MD+2MD^{1-s}+\frac{2MD^{2-s}}{\sigma^2}+DMC_\rho,\end{aligned}$$

so that

$$\begin{aligned}
\|\mathrm{I}_a\|^2_{\mathcal{H}_K^n} &\le nC\alpha^{2t}\max_i\Big\{2\kappa^2 MD\sup_{\mathbf{x}}|h^i(\mathbf{x})| \\
&\qquad + \Big|\iint_X h^i(\mathbf{x})K(\mathbf{x},\mathbf{r})|w^{\sigma}_{\mathbf{r},\boldsymbol{\tau}}(\boldsymbol{\tau}^i-\mathbf{r}^i)(f_\rho(\boldsymbol{\tau})-f_\rho(\mathbf{r}))|_{\boldsymbol{\tau},C^s(X)}d\rho_X(\mathbf{x})\,d\rho_X(\mathbf{r})\Big|\Big\} \\
&\le nC\alpha^{2t}\Big\{2\kappa^2MD+\kappa^2\Big(2MD^{1-s}+\frac{2MD^{2-s}}{\sigma^2}+DMC_\rho\Big)\Big\}\sup_{\mathbf{x}}|h^i(\mathbf{x})| \\
&\le n(C\alpha^{2t})^2\kappa^2\Big\{2MD+2MD^{1-s}+\frac{2MD^{2-s}}{\sigma^2}+DMC_\rho\Big\}^2.
\end{aligned}$$

Similarly, we can get the upper bound

$$\|\mathrm{I}_b\|^2_{\mathcal{H}_K^n}\le n(C\alpha^{2t-1})^2(\kappa^2+2\kappa_s+|K|_{C^s(X\times X)})\Big\{2MD+2MD^{1-s}+\frac{2MD^{2-s}}{\sigma^2}+DMC_\rho\Big\}^2.$$

Then $\|\mathbf{f}_{\rho,\sigma}-\mathbf{f}_{\rho^{(2t)},\sigma}\|_{\mathcal{H}_K^n}\le\sqrt{2(\|\mathrm{I}_a\|^2_{\mathcal{H}_K^n}+\|\mathrm{I}_b\|^2_{\mathcal{H}_K^n})}\le C_1\alpha^{2t-1}/\sigma$, with C_1 being independent of α,σ.

Next, for $(L_{K,\rho^{(2t)}}-L_{K,\rho})\mathbf{f}_{\lambda,\rho^{(2t)}}$, we give the analysis of its general case, that is, $L_{K,\rho^{(2t)}}\mathbf{f}-L_{K,\rho}\mathbf{f}$ with $\mathbf{f}\in\mathcal{H}_K^n$. It can be rewritten as

$$\begin{aligned}
L_{K,\rho^{(2t)}}\mathbf{f}-L_{K,\rho}\mathbf{f} &= \int_X\int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^T\mathbf{f}(\mathbf{x})K_{\mathbf{x}}\,d\rho_X^{(2t)}(\mathbf{u})\,d(\rho_X^{(2t-1)}-\rho_X)(\mathbf{x}) \\
&\quad + \int_X\int_X w^{\sigma}_{\mathbf{x},\mathbf{u}}(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^T\mathbf{f}(\mathbf{x})K_{\mathbf{x}}\,d(\rho_X^{(2t)}-\rho_X)(\mathbf{u})\,d\rho_X(\mathbf{x}) \\
&= \mathrm{II}_a+\mathrm{II}_b.
\end{aligned}$$

Using similar techniques to those in the proof of the first part,

$$\begin{aligned}
\|\mathrm{II}_a\|^2_{\mathcal{H}_K^n} &\le n(C\alpha^{2t-1})^2[(D+D^{3-s}/\sigma+2D^{2-s}/\sigma^2)|\mathbf{f}|_{C^s(X)}+D^2|\mathbf{f}|_\infty]^2 \\
&\qquad\times(\kappa^2+2\kappa_s+|K|_{C^s(X\times X)})^2, \\
\|\mathrm{II}_b\|^2_{\mathcal{H}_K^n} &\le n(C\alpha^{2t}\kappa)^2[(D+D^{3-s}/\sigma+2D^{2-s}/\sigma^2)|\mathbf{f}|_{C^s(X)}+D^2|\mathbf{f}|_\infty]^2.
\end{aligned}$$

Therefore $\sqrt{2(\|\mathrm{II}_a\|^2_{\mathcal{H}_K^n}+\|\mathrm{II}_b\|^2_{\mathcal{H}_K^n})}\le C_2\alpha^{2t-1}\|\mathbf{f}\|_{C^s(X)}/\sigma$, with C_2 being independent of α,σ.

From (2.4),

$$\lambda\|\mathbf{f}_{\lambda,\rho^{(2t)}}\|^2_{\mathcal{H}_K^n}\le\int_Z\int_Z w^{\sigma}_{\mathbf{x},\mathbf{u}}(y-v)^2\,d\rho^{(2t-1)}(\mathbf{x},y)\,d\rho^{(2t)}(\mathbf{u},v)\le D^2, \tag{3.1}$$

which yields

$$\|\mathbf{f}_{\lambda,\rho^{(2t)}}\|_{C^s(X)}=\sum_{j=1}^n\|f^j_{\lambda,\rho^{(2t)}}\|_{C^s(X)}\le(\kappa+\kappa_s)\sum_{j=1}^n\|f^j_{\lambda,\rho^{(2t)}}\|_K\le(\kappa+\kappa_s)D\sqrt{\frac{n}{\lambda}}.$$

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

Combining the above results,

$$\|\mathbf{f}_{\lambda,\rho} - \mathbf{f}_{\lambda,\rho^{(2t)}}\|_{\mathcal{H}_K^n} \le \frac{1}{\lambda}\bigg\{C_1\frac{\alpha^{2t-1}}{\sigma} + C_2\frac{\alpha^{2t-1}}{\sigma}(\kappa+\kappa_s)D\sqrt{\frac{n}{\lambda}}\bigg\} \le C'\alpha^{2t-1}\lambda^{-3/2}\sigma^{-1}.$$

This completes the proof. □

To prove the convergence rate of algorithm (2.3), we need to find an upper bound for a difference caused by the change of the regularization parameter from λ_i to λ_{i+1} in (2.5).

Proposition 3.2. *Let* $\lambda_i = \lambda_1 i^{-\beta}$ *with* $0 < \lambda_1, \beta < 1$ *and* $\mu \in (C^s(X))^*$. *Then*

$$\|\mathbf{f}_{\lambda_i,\mu} - \mathbf{f}_{\lambda_{i+1},\mu}\|_{\mathcal{H}_K^n} \le \frac{4\kappa MD\beta}{\lambda_1} i^{\beta-1}.$$

Proof. Using the notation in (2.5),

$$\begin{aligned}\mathbf{f}_{\lambda_i,\mu} - \mathbf{f}_{\lambda_{i+1},\mu} &= (L_{K,\mu} + \lambda_i I)^{-1}(\mathbf{f}_{\mu,\sigma} - (L_{K,\mu} + \lambda_{i+1} I)\mathbf{f}_{\lambda_{i+1},\mu} + (\lambda_{i+1} I - \lambda_i I)\mathbf{f}_{\lambda_{i+1},\mu})\\ &= (\lambda_{i+1} I - \lambda_i I)(L_{K,\mu} + \lambda_i I)^{-1}(L_{K,\mu} + \lambda_{i+1} I)^{-1}\mathbf{f}_{\mu,\sigma},\end{aligned}$$

so that $\|\mathbf{f}_{\lambda_i,\mu} - \mathbf{f}_{\lambda_{i+1},\mu}\|_{\mathcal{H}_K^n} \le ((\lambda_i - \lambda_{i+1})/\lambda_i\lambda_{i+1})\|\mathbf{f}_{\mu,\sigma}\|_{\mathcal{H}_K^n}$. Since $\lambda_i - \lambda_{i+1} = \lambda_1\beta\xi^{-\beta-1}$ for some $\xi \in (i, i+1)$, this yields

$$\|\mathbf{f}_{\lambda_i,\mu} - \mathbf{f}_{\lambda_{i+1},\mu}\|_{\mathcal{H}_K^n} \le \frac{\lambda_1\beta i^{-\beta-1}}{\lambda_1 i^{-\beta}\lambda_1(i+1)^{-\beta}}\|\mathbf{f}_{\mu,\sigma}\|_{\mathcal{H}_K^n} \le \frac{2\beta}{\lambda_1}i^{\beta-1}\|\mathbf{f}_{\mu,\sigma}\|_{\mathcal{H}_K^n} \le \frac{2\beta}{\lambda_1}i^{\beta-1}(2\kappa MD).$$

□

We also need the following revised McDiarmid–Bernstein-type probability inequality that was originally proposed by Mukherjee and Zhou [7].

Lemma 3.3. *Let* $S = \{\mathbf{z}_i\}_{i=1}^m$ *be independently drawn from probability distributions* $\{\rho^{(i)}\}_{i=1}^m$, *respectively, let* $(H, \|\cdot\|)$ *be a Hilbert space and let* $F : Z^m \to H$ *be measurable. If there is* $\widetilde{M} \ge 0$ *such that* $\|F(S) - \mathbb{E}_{\mathbf{z}_i}(F(S))\| \le \widetilde{M}$ *for each* $1 \le i \le m$ *and almost every* $S \in Z^m$, *then, for any* $\varepsilon > 0$,

$$\mathbb{P}_{S\in Z^m}\{\|F(S) - \mathbb{E}_S(F(S))\| \ge \varepsilon\} \le 2\exp\bigg\{-\frac{\varepsilon^2}{2(\widetilde{M}\varepsilon + m\widetilde{M}^2)}\bigg\}.$$

The proof of this lemma is similar to a result of Mukherjee and Zhou [7, Proposition 13], and hence is omitted here. Now we present our second main result.

Theorem 3.4. *Assume that the true gradient of the regression function* f_ρ *is* $\nabla f_\rho \in \mathcal{H}_K^n$. *Taking* $\lambda_t = \lambda_1 t^{-\beta}$, $\eta_t = \eta_1 t^{-\theta}$ *and* $\sigma = t^{-3/2\beta}(t \in \mathbb{N})$ *with* $0 < 2\beta < \theta < 1/2$,

$$\mathbb{E}_{z_1,\ldots,z_{2t}}(\|\mathbf{f}_{t+1} - \nabla f_\rho\|_{\mathcal{H}_K^n}) \le \widetilde{C}\Big(\frac{1}{t}\Big)^{\min\{\beta/2,(\theta/2)-\beta\}}.$$

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

Proof. From (2.6), first we need to estimate the sample error. Denote $W_{t+1} = \mathbf{f}_{t+1} - \mathbf{f}_{\lambda_t,\rho^{(2t)}}$ with $\mathbf{f}_{\lambda_0,\rho^{(0)}} = 0$. Then

$$\begin{aligned} W_{t+1} &= \mathbf{f}_t - \mathbf{f}_{\lambda_t,\rho^{(2t)}} - \eta_t\{S^T_{\mathbf{x}_{2t-1}} D_{\mathbf{x}_{2t}} S_{\mathbf{x}_{2t-1}}(\mathbf{f}_t) - S^T_{\mathbf{x}_{2t-1}} Y_{\mathbf{x}_{2t}} + \lambda_t \mathbf{f}_t\} \\ &= \mathbf{f}_t - \mathbf{f}_{\lambda_t,\rho^{(2t)}} - \eta_t\{S^T_{\mathbf{x}_{2t-1}} D_{\mathbf{x}_{2t}} S_{\mathbf{x}_{2t-1}}(\mathbf{f}_t - \mathbf{f}_{\lambda_t,\rho^{(2t)}}) + S^T_{\mathbf{x}_{2t-1}} D_{\mathbf{x}_{2t}} S_{\mathbf{x}_{2t-1}}(\mathbf{f}_{\lambda_t,\rho^{(2t)}}) \\ &\quad - S^T_{\mathbf{x}_{2t-1}} Y_{\mathbf{x}_{2t}} + \lambda_t \mathbf{f}_t\}. \end{aligned}$$

Let $A_t = (1 - \eta_t\lambda_t)I - \eta_t S^T_{\mathbf{x}_{2t-1}} D_{\mathbf{x}_{2t}} S_{\mathbf{x}_{2t-1}}$, $\boldsymbol{\xi}_t = S^T_{\mathbf{x}_{2t-1}} D_{\mathbf{x}_{2t}} S_{\mathbf{x}_{2t-1}}(\mathbf{f}_{\lambda_t,\rho^{(2t)}}) - S^T_{\mathbf{x}_{2t-1}} Y_{\mathbf{x}_{2t}} + \mathbf{f}_{\rho^{(2t)},\sigma} - L_{K,\rho^{(2t)}}(\mathbf{f}_{\lambda_t,\rho^{(2t)}})$ and denote $\Pi^t_{j=t+1} A_j = I$. Then, by iteration,

$$W_{t+1} = \sum_{i=1}^{t} \prod_{j=i}^{t} A_j(\mathbf{f}_{\lambda_{i-1},\rho^{(2i-2)}} - \mathbf{f}_{\lambda_i,\rho^{(2i)}}) - \sum_{i=1}^{t} \prod_{j=i+1}^{t} A_j \eta_i \boldsymbol{\xi}_i. \tag{3.2}$$

In the following, we will analyse the two terms on the right-hand side of equation (3.2). The operator $\eta_j\lambda_j I + \eta_j S^T_{x_{2j-1}} D_{x_{2j}} S_{x_{2j-1}}$ is positive and bounded by $(\eta_j\lambda_j + \eta_j\kappa D^2)I$. So, for $j \geq t_0$, where t_0 is the smallest integer greater than $(\eta_1\lambda_1 + \eta_1\kappa D^2)^{1/\theta}$, the operator $A_j : \mathcal{H}^n_K \to \mathcal{H}^n_K$ is positive and bounded by $(1 - \eta_j\lambda_j)I$, and hence $\|A_j\|_{\mathcal{H}^n_K \to \mathcal{H}^n_K} \leq \exp\{-\eta_j\lambda_j\}$. Since $\|A_j\|_{\mathcal{H}^n_K \to \mathcal{H}^n_K} \leq 1 + \eta_j\lambda_j + \eta_j\kappa D^2$ for $j < t_0$,

$$\begin{aligned} \Big\|\prod_{j=i}^{t} A_j\Big\|_{\mathcal{H}^n_K \to \mathcal{H}^n_K} &\leq (1 + \eta_1\lambda_1 + \eta_1\kappa D^2)^{(\eta_1\lambda_1+\eta_1\kappa D^2)^{1/\theta}} \exp\Big\{-\eta_1\lambda_1 \sum_{j=i}^{t} j^{-\beta-\theta}\Big\} \\ &= C_0 \exp\Big\{-\eta_1\lambda_1 \sum_{j=i}^{t} j^{-\beta-\theta}\Big\}. \end{aligned} \tag{3.3}$$

From Theorem 3.1 and Proposition 3.2,

$$\begin{aligned} \|\mathbf{f}_{\lambda_{i-1},\rho^{(2i-2)}} - \mathbf{f}_{\lambda_i,\rho^{(2i)}}\|_{\mathcal{H}^n_K} &\leq \|\mathbf{f}_{\lambda_{i-1},\rho^{(2i-2)}} - \mathbf{f}_{\lambda_{i-1},\rho}\|_{\mathcal{H}^n_K} + \|\mathbf{f}_{\lambda_i,\rho} - \mathbf{f}_{\lambda_i,\rho^{(2i)}}\|_{\mathcal{H}^n_K} + \|\mathbf{f}_{\lambda_{i-1},\rho} - \mathbf{f}_{\lambda_i,\rho}\|_{\mathcal{H}^n_K} \\ &\leq C_1'\alpha^{2i-3}(i-1)^{(3/2)\beta} + C_2'\alpha^{2i-1}i^{(3/2)\beta} + C_3'(i-1)^{\beta-1}. \end{aligned} \tag{3.4}$$

By Lemma 2(1) of Smale and Zhou [10], since $\alpha^{2i} = \exp\{-2i\ln(1/\alpha)\} \leq (b/2e\ln(1/\alpha))^b i^{-b}$, if we choose $b = 3\beta/2$, the term with $\alpha^{2i}i^{3\beta/2}$ in (3.4) is dominated by the polynomial term $(i-1)^{\beta-1}$. Equation (3.3) and [10, Lemma 2(2)], together with $\nu = \eta_1\lambda_1, p_1 = \beta + \theta, p_2 = 1 - \beta$, give

$$\Big\|\sum_{i=1}^{t} \prod_{j=i}^{t} A_j(\mathbf{f}_{\lambda_{i-1},\rho^{(2i-2)}} - \mathbf{f}_{\lambda_i,\rho^{(2i)}})\Big\|_{\mathcal{H}^n_K} \leq C'' C_{\nu,p_1,p_2} t^{\theta+2\beta-1}. \tag{3.5}$$

Now we estimate the second term in the right-hand side of equation (3.2). Write

$$\begin{aligned} \Big\|\sum_{i=1}^{t} \prod_{j=i+1}^{t} A_j\eta_i\boldsymbol{\xi}_i\Big\|^2_{\mathcal{H}^n_K} &= \sum_{i=1}^{t-1}\sum_{l=1}^{t-1}\Big\langle \prod_{j=i+1}^{t} A_j\eta_i\boldsymbol{\xi}_i, \prod_{p=l+1}^{t} A_p\eta_l\boldsymbol{\xi}_l\Big\rangle_{\mathcal{H}^n_K} \\ &\quad + \Big\langle \sum_{i=1}^{t-1}\prod_{j=i+1}^{t} A_j\eta_i\boldsymbol{\xi}_i, \eta_t\boldsymbol{\xi}_t\Big\rangle_{\mathcal{H}^n_K} + \|\eta_t\boldsymbol{\xi}_t\|^2_{\mathcal{H}^n_K}, \end{aligned}$$

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

and denote $\widetilde{\mathbf{z}}_i = \{\mathbf{z}_{2i-1}, \mathbf{z}_{2i}\}$. Notice that $\boldsymbol{\xi}_i$ depends on $\mathbf{z}_{2i-1}, \mathbf{z}_{2i}$ and $\mathbb{E}_{\widetilde{\mathbf{z}}_i}(\boldsymbol{\xi}_i) = 0$, while, for $i < t$, $\prod_{j=i+1}^{t} A_j$ depends on $\mathbf{z}_{2t}, \mathbf{z}_{2t-1}, \ldots, \mathbf{z}_{2i+1}$, which yields

$$\mathbb{E}_{\widetilde{\mathbf{z}}_i|\mathbf{z}_{2t},\mathbf{z}_{2t-1},\ldots,\mathbf{z}_{2i+1}}\bigg(\prod_{j=i+1}^{t} A_j \eta_i \boldsymbol{\xi}_i\bigg) = 0.$$

It follows that, for $l > i + 1$, the expected value

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_{2t}}\bigg(\bigg\langle \prod_{j=i+1}^{t} A_j \eta_i \boldsymbol{\xi}_i, \prod_{p=l+1}^{t} A_p \eta_l \boldsymbol{\xi}_l\bigg\rangle_{\mathcal{H}_K^n}\bigg)\\ &\quad= \mathbb{E}_{\mathbf{z}_{2t},\mathbf{z}_{2t-1},\ldots,\mathbf{z}_{2i+1}}\bigg\langle \mathbb{E}_{\widetilde{\mathbf{z}}_i|\mathbf{z}_{2t},\mathbf{z}_{2t-1},\ldots,\mathbf{z}_{2i+1}} \prod_{j=i+1}^{t} A_j \eta_i \boldsymbol{\xi}_i, \prod_{p=l+1}^{t} A_p \eta_l \boldsymbol{\xi}_l\bigg\rangle_{\mathcal{H}_K^n}\\ &\quad= 0.\end{aligned}$$

Therefore

$$\mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_{2t}}\bigg(\bigg\|\sum_{i=1}^{t}\prod_{j=i+1}^{t} A_j \eta_i \boldsymbol{\xi}_i\bigg\|_{\mathcal{H}_K^n}^2\bigg) = \sum_{i=1}^{t-1} \mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_{2t}}\bigg(\bigg\|\prod_{j=i+1}^{t} A_j \eta_i \boldsymbol{\xi}_i\bigg\|_{\mathcal{H}_K^n}^2\bigg) + \mathbb{E}_{\widetilde{\mathbf{z}}_t}\|\eta_t \boldsymbol{\xi}_t\|_{\mathcal{H}_K^n}^2. \quad (3.6)$$

Let $H = \mathcal{H}_K^n$. Considering the function $F : Z^2 \to H$ given by

$$F(\widetilde{\mathbf{z}}_t) = S_{\mathbf{x}_{2t-1}}^T D_{\mathbf{x}_{2t}} S_{\mathbf{x}_{2t-1}}(\mathbf{f}_{\lambda_t,\rho^{(2t)}}) - S_{\mathbf{x}_{2t-1}}^T Y_{\mathbf{x}_{2t}}, \quad \boldsymbol{\xi}_t = F(\widetilde{\mathbf{z}}_t) - \mathbb{E}_{\widetilde{\mathbf{z}}_t}[F(\widetilde{\mathbf{z}}_t)].$$

Since

$$\mathbb{E}_{\mathbf{z}_{2t-1}}[F(\widetilde{\mathbf{z}}_t)] = \int_X w_{\mathbf{x},\mathbf{x}_{2t}}^{\sigma}(f_\rho(\mathbf{x}) - y_{2t} + \mathbf{f}_{\lambda_t,\rho^{(2t)}}(\mathbf{x}) \cdot (\mathbf{x}_{2t} - \mathbf{x}))(\mathbf{x}_{2t} - \mathbf{x}) K_{\mathbf{x}}\, d\rho_X^{(2t-1)}(\mathbf{x}),$$

by the reproducing property of the RKHS and inequality (3.1),

$$\|F(\widetilde{\mathbf{z}}_t) - \mathbb{E}_{\mathbf{z}_{2t-1}}[F(\widetilde{\mathbf{z}}_t)]\|_{\mathcal{H}_K^n} \le 2D\kappa(2M + D\kappa\|\mathbf{f}_{\lambda_t,\rho^{(2t)}}\|_{\mathcal{H}_K^n}) = M_t.$$

A similar result can be obtained for $\|F(\widetilde{\mathbf{z}}_t) - \mathbb{E}_{\mathbf{z}_{2t}}[F(\widetilde{\mathbf{z}}_t)]\|_{\mathcal{H}_K^n} \le M_t$.

Then, by applying Lemma 3.3 with $\widetilde{M} = M_t$,

$$\mathbb{P}\{\|\boldsymbol{\xi}_t\|_{\mathcal{H}_K^n} \ge \varepsilon\} \le 2\exp\bigg\{-\frac{\varepsilon^2}{2(M_t\varepsilon + 2M_t^2)}\bigg\}$$

for any $0 < \varepsilon < 1$. Combining this with

$$\mathbb{E}_{\widetilde{\mathbf{z}}_t}[\|\boldsymbol{\xi}_t\|_{\mathcal{H}_K^n}^2] = \int_0^\infty \mathbb{P}\{\|\boldsymbol{\xi}_t\|_{\mathcal{H}_K^n}^2 \ge \varepsilon\} d\varepsilon = \int_0^\infty \mathbb{P}\{\|\boldsymbol{\xi}_t\|_{\mathcal{H}_K^n} \ge \sqrt{\varepsilon}\} d\varepsilon,$$

we see that, for any $u > 0$, $\mathbb{E}_{\widetilde{\mathbf{z}}_t}[\|\boldsymbol{\xi}_t\|_{\mathcal{H}_K^n}^2] \le u + 2\int_u^\infty \exp\{-\varepsilon/2M_t(\sqrt{\varepsilon} + 2M_t)\} d\varepsilon$. If $u \ge (2M_t)^2$,

$$\begin{aligned}\int_u^\infty \exp\bigg\{-\frac{\varepsilon}{2M_t(\sqrt{\varepsilon} + 2M_t)}\bigg\} d\varepsilon &\le \int_u^\infty \exp\bigg\{-\frac{\sqrt{\varepsilon}}{4M_t}\bigg\} d\varepsilon\\ &= 8M_t(\sqrt{u} + 4M_t)\exp\bigg\{-\frac{\sqrt{u}}{4M_t}\bigg\}.\end{aligned}$$

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

Let $f(u) = u + 16M_t(\sqrt{u} + 4M_t)\exp\{-\sqrt{u}/4M_t\}$. It is easy to prove that the minimizer of $f(u)$ is $u_0 = (4M_t \ln 2)^2$ and that $f(u_0) = (4M_t \ln 2)^2 + 32M_t^2(\ln 2 + 1)$. Thus $\mathbb{E}_{\widetilde{\mathbf{z}}_t}[\|\boldsymbol{\xi}_t\|^2_{\mathcal{H}_K^n}] \leq (4M_t \ln 2)^2 + 32M_t^2(\ln 2 + 1) \leq 48M_t^2$. Substituting (3.1) into (3.6),

$$\mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_{2t}}\bigg(\bigg\|\sum_{i=1}^{t}\prod_{j=i+1}^{t} A_j\eta_i\boldsymbol{\xi}_i\bigg\|^2_{\mathcal{H}_K^n}\bigg) \leq \sum_{i=1}^{t-1} c''' i^{-2\theta+\beta}\exp\bigg\{-2\eta_1\lambda_1\sum_{j=i+1}^{t} j^{-\beta-\theta}\bigg\}.$$

Using [10, Lemma 2(2)], together with equations (3.1), (3.5) and (3.3), we can find an upper bound for the sample error as

$$\mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_{2t}}(\|W_{t+1}\|_{\mathcal{H}_K^n}) \leq \widetilde{C}_1 t^{\theta+2\beta-1} + \widetilde{C}_2 t^{\beta-(\theta/2)}. \tag{3.7}$$

For the approximation error in (2.6), we use [7, Proposition 9], namely,

$$\|\mathbf{f}_{\lambda_t,\rho} - \nabla f_\rho\|_{\mathcal{H}_K^n} \leq \widetilde{C}_3\bigg\{\frac{\sigma}{\lambda_t} + \sqrt{\lambda_t}\bigg\}.$$

Combining this with (3.7) and Theorem 3.1,

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_{2t}}(\|\mathbf{f}_{t+1} - \nabla f_\rho\|_{\mathcal{H}_K^n}) \leq \widetilde{C}_1 t^{\theta+2\beta-1} + \widetilde{C}_2 t^{\beta-(\theta/2)} + \widetilde{C}_0\frac{\alpha^{2t-1}t^{3\beta/2}}{\lambda_1^{3/2}\sigma}\\ + \widetilde{C}_3\bigg\{\frac{\sigma}{\lambda_1}t^{\beta} + \sqrt{\lambda_1}t^{-\beta/2}\bigg\}.\end{aligned}$$

Using [10, Lemma 2(2)], we derive that $(\alpha^{2t-1}t^{3\beta/2})/\lambda_1^{3/2}\sigma$ can be dominated by $1/\sigma t^{2\beta}$, and thus, under the condition of the theorem, we have proved the desired result. □

4. Conclusions

In this work, an online gradient-learning algorithm is described that can provide information of variable selection and coordinate covariance estimation for nonidentical data. Under certain conditions, we show that the gradient derived by the algorithm is an approximation of the true gradient of the regression function. Interesting areas for future directions include using the proposed algorithm in network security for recognition of malicious online attacks or for other related research areas, and to improve the learning rate through choosing parameters adaptively.

Acknowledgements

The author would like to thank Professor Dianhui Wang for his help in technical writing and academic training during her stay at La Trobe University as a visiting scholar from May 2015 to April 2016. This work was supported in part by Zhejiang Provincial Natural Science Foundation under grant LQ12A01016, National Natural Science Foundation of China under grant 11201420, China Scholarship Council Fund and Zhejiang Provincial Key Research Base for Humanities and Social Science Research (Statistics).

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press

References

[1] N. Aronszajn, “Theory of reproducing kernels”, *Trans. Amer. Math. Soc.* **68** (1950) 337–404; doi:10.2307/1990404.

[2] P. Bickel and E. Levina, “Covariance regularization by thresholding”, *Ann. Statist.* **36** (2008) 2577–2604; doi:10.1214/08-AOS600.

[3] T. Cai and W. Liu, “Adaptive thresholding for sparse covariance matrix estimation”, *J. Amer. Statist. Assoc.* **106** (2011) 672–684; doi:10.1198/jasa.2011.tm10560.

[4] I. Guyon and A. Ellsseeff, “An introduction to variable and feature selection”, *J. Mach. Learn. Res.* **3** (2003) 1157–1182; doi:10.1162/153244303322753616.

[5] T. Koski, *Hidden Markov models for bioinformatics* (Springer, Netherlands, 2001).

[6] P. D. Lax, *Functional analysis* (John Wiley & Sons, New York, 2002).

[7] S. Mukherjee and D. X. Zhou, “Learning coordinate covariances via gradients”, *J. Mach. Learn. Res.* **7** (2006) 519–549;
http://www.jmlr.org/papers/volume7/mukherjee06a/mukherjee06a.pdf.

[8] A. E. Raftery, D. Madigan and J. A. Hoeting, “An introduction to variable and feature selection”, *J. Amer. Statist. Assoc.* **92** (1998) 179–191; doi:10.1080/01621459.1997.10473615.

[9] C. Robinson, *Dynamical systems: stability, symbolic dynamics, and chaos* (CRC Press, New York, 1998).

[10] S. Smale and D.-X. Zhou, “Online learning with Markov sampling”, *Anal. Appl.* **7** (2009) 87–113; doi:10.1142/S0219530509001293.

[11] I. Steinwart, D. Hush and C. Scovel, “Learning from dependent observations”, *J. Multivariate Anal.* **100** (2009) 175–194; doi:10.1016/j.jmva.2008.04.001.

[12] R. Tibshirani, “Regression shrinkage and selection via the lasso”, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** (1996) 267–288; doi:10.2307/2346178.

[13] G. B. Ye and X. Xie, “Learning sparse gradients for variable selection and dimension reduction”, *Mach. Learn.* **87** (2012) 303–355; doi:10.1007/s10994-012-5284-9.

[14] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net”, *J. R. Stat. Soc. Ser.* B*, Stat. Methodol.* **67** (2005) 301–320; doi:10.1111/j.1467-9868.2005.00503.x.

https://doi.org/10.1017/S1446181116000328 Published online by Cambridge University Press