MI Preprint Series

Mathematics for Industry
Kyushu University

Moment convergence of regularized least-squares estimator for linear regression model

Yusuke Shimizu

MI 2014-13

(Received December 6, 2014)

Institute of Mathematics for Industry
Graduate School of Mathematics
Kyushu University
Fukuoka, JAPAN

Moment convergence of regularized least-squares estimator for linear regression model*

Yusuke Shimizu[†]

Abstract

In this paper we study the uniform tail-probability estimates of a regularized least-squares estimator for the linear regression model, by making use of the polynomial type large deviation inequality for the associated statistical random fields, which may not be locally asymptotically quadratic. Our results provide a measure of rate of consistency in variable selection in sparse estimation, which in particular enable us to verify various arguments requiring convergence of moments of estimator-dependent statistics, such as the expected maximum-likelihood for AIC-type and many other moment based model assessment procedure including the C_p-type.

Keywords Moment convergence · Regularized least-squares estimation · Large deviation inequality · Sparse estimation

1 Introduction

Assume that we have a sample $\{(X_i, Y_i)\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ and $X_i = (X_{i,1}, \dots, X_{i,p}) \in \mathbb{R}^p$, obeying the linear regression model:

$$Y_i = \theta_0^\top X_i + \epsilon_i, \quad i = 1, \dots, n, \tag{1.1}$$

where θ_0 is a p-dimensional true value of parameter contained in the interior of a compact parameter space $\Theta \subset \mathbb{R}^p$ and $\epsilon_1, \epsilon_2, \dots$ represent noises. Through this paper, the number of variables p is fixed. Though not essential, we suppose that the covariate X is non-random; usually $X_1, \cdots, X_n$ are standardized from the beginning, but for brevity we omit the dependence of X_i and Y_i on n from the notation. In this paper we deal with the situation

$$\theta_0 = (z_0, \rho_0) = (z_{0,1}, \dots, z_{0,p_0}, \rho_{0,1}, \dots, \rho_{0,p_1}),$$

where $z_{0,k} = 0$ and $\rho_{0,l} \neq 0$ for any $k \in \{1, \dots, p_0\}$ and $l \in \{1, \dots, p_1\}$; divide the compact parameter space $\Theta = \Theta_0 \times \Theta_1 \subset \mathbb{R}^{p_0} \times \mathbb{R}^{p_1}$ such that $z_0 = 0 \in \Theta_0$ and $\rho_0 \in \Theta_1$. We can rewrite the linear regression model (1.1) to

$$Y_i = z_0^\top X_i^{(z)} + \rho_0^\top X_i^{(\rho)} + \epsilon_i, \quad i = 1, \dots, n, \tag{1.2}$$

*This version: December 6, 2014

[†]Graduate school of Mathematics, Kyushu University. 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan. Email: y-shimizu@math.kyushu-u.ac.jp

where $X_i^{(z)} := (X_{i,1}, \dots, X_{i,p_0})$ and $X_i^{(\rho)} := (X_{i,p_0+1}, \dots, X_{i,p_0+p_1})$, representing irrelevant and relevant covariate vectors, respectively. Then we define the regularized least-squares estimator (regularized-LSE) $\hat{\theta}_n = (\hat{z}_n, \hat{\rho}_n)$ as the minimizer of the contrast function

$$Z_n(\theta) = Z_n(z, \rho) := \sum_{i=1}^{n} (Y_i - z^\top X_i^{(z)} - \rho^\top X_i^{(\rho)})^2 + \sum_{j=1}^{p} \mathfrak{p}_n(\theta_j) \tag{1.3}$$

over Θ, where $\mathfrak{p}_n(\cdot)$ is a non-random non-negative function such that $\mathfrak{p}_n(0) = 0$. There is a huge literature on the sparse linear regression via regularization, where the estimator $\hat{z}_n$ of $z_0 = 0$ satisfies the *sparse consistency* $P(\hat{z}_n = 0) \to 1$ as $n \to \infty$, which implies that $R_n \hat{z}_n = o_p(1)$ for arbitrary $R_n \to \infty$, while $\sqrt{n}(\hat{\rho}_n - \rho_0)$ has a non-trivial asymptotic law; e.g. sparse-bridge (Radchenko [6]), the smoothly clipped absolute deviation (SCAD; Fan and Li [2]) and the seamless-L_0 regularization (Dicker et al. [1]). In Section 3, we will refer some asymptotic behaviors of these regularized estimators.

We will prove the moment convergence of the scaled estimator

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \big(\sqrt{n}\hat{z}_n, \sqrt{n}(\hat{\rho}_n - \rho_0)\big).$$

Let us mention some basic facts concerning the parametric M-estimation. Given a statistical model indexed by a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^p$, we typically estimate the true parameter value $\theta_0 \in \Theta$ by a minimum point $\hat{\theta}_n$ of an appropriate continuous contrast function $Z_n : \Theta \to \mathbb{R}$. In order to assess the asymptotic performance of $\hat{\theta}_n$ quantitatively, when $\sqrt{n}$-consistency is concerned, we look at the statistical random fields

$$\mathbb{M}_n(w; \theta_0) := Z_n\Big(\theta_0 + \frac{w}{\sqrt{n}}\Big) - Z_n(\theta_0), \tag{1.4}$$

where $w \in \mathbb{R}^p$. As is well-known, the weak convergence of $\mathbb{M}_n$ to some $\mathbb{M}_0$ over compact sets, the identifiability condition on $\mathbb{M}_0$, and the tightness of the scaled estimator $\hat{w}_n := \sqrt{n}(\hat{\theta}_n - \theta_0)$ make the "argmin" functional continuous for $\mathbb{M}_n$: $\hat{w}_n \in \operatorname{argmin} \mathbb{M}_n \xrightarrow{\mathcal{L}} \operatorname{argmin} \mathbb{M}_0$. See e.g., van der Vaart [11]. Further, when concerned with moments of $\hat{w}_n$-dependent statistics such as the mean square error, more than the weak convergence is required. Then the *polynomial type large deviation inequality (PLDI)* of Yoshida [12], which estimates the tail of $\mathcal{L}(\hat{w}_n)$ in such a way that

$$\sup_{r>0} \sup_{n>0} r^L P(|\hat{w}_n| \geq r) < \infty \tag{1.5}$$

for a given $L > 0$, plays an important role. When $\hat{w}_n \xrightarrow{\mathcal{L}} \hat{w}_0$ for a random variable $\hat{w}_0$, the moment convergence

$$E[|\hat{w}_n|^q] \to E[|\hat{w}_0|^q], \quad q > 0 \tag{1.6}$$

holds if there exists a $q' > q$ such that $\sup_{n>0} E[|\hat{w}_n|^{q'}] < \infty$. Assume that the PLDI (1.5) holds for some $L > q'$. Then we obtain

$$\sup_{n>0} E[|\hat{w}_n|^{q'}] = \sup_{n>0} \int_0^\infty P(|\hat{w}_n|^{q'} > s) ds < \infty.$$

As the results, the moment convergence (1.6) holds if we can ensure the PLDI (1.5) for some $L > q$.

The main purpose of this paper is to derive the moment convergence of $\hat{w}_n = (\sqrt{n}\hat{z}_n, \sqrt{n}(\hat{\rho}_n - \rho_0))$: if we have the weak convergence $\big(\sqrt{n}\hat{z}_n, \sqrt{n}(\hat{\rho}_n - \rho_0)\big) \xrightarrow{\mathcal{L}} (\hat{u}_0, \hat{v}_0)$ for some random vector $(\hat{u}_0, \hat{v}_0) =: \hat{w}_0$, then for every continuous $f : \mathbb{R}^p \to \mathbb{R}$ of at most polynomial growth we have

$$E[f(\hat{w}_n)] \to E[f(\hat{w}_0)], \tag{1.7}$$

through the PLDI (1.5): for any $L > 0$ there exists a constant $c_L > 0$ for which

$$\sup_{n>0} P(|\hat{w}_n| \geq r) < \frac{c_L}{r^L}, \quad r > 0. \tag{1.8}$$

Let us briefly remark the importance of convergence of moments: asymptotic behavior of expected values of statistics depending on estimators. The PLDI for statistical random fields associated with stochastic process have been studied and applied for example to the information criteria in model selection, the higher-order statistics, as well as the moment convergence for Gaussian quasi-likelihood and Bayes estimators of diffusion processes; see Uchida and Yoshida [9, 10], Sakamoto and Yoshida [8], Yoshida [12] and the references therein for details. See also Masuda [4] for the PLDI associated with the Gaussian quasi-likelihood estimation of a Lévy driven stochastic differential equation, as well as for more related references.

It has been known that the PLDI can be proved under modest conditions when $\mathbb{M}_n$ is smooth and well-integrable, and further admits a *partially locally asymptotically quadratic (PLAQ)* structure (see (1.9) below), which is satisfied for many situations. However in the regularized estimations, the key PLAQ structure may break down; it may happen that $r_n(w; \theta_0)$ diverges in probability. As a matter of fact, most of the existing sparse-estimation procedures belong to this type of asymptotics. Therefore, our results provide a theoretically deeper understanding on the recently highlighted sparse estimation.

This paper is organized as follows. In Section 2, we will derive the PLDI for the regularized-LSE of the linear regression model (1.2). We will look at the PLDI for the random fields only associated with the zero parameter z in Section 2.2. In Section 3, we will give some examples of the regularization term in the contrast function (1.3).

For convenience of reference, we end this section with stating Yoshida [12, Theorem 1, Theorem 3(a)], which will play an essential role in our study. We need to introduce some notation. Given a set $K \subset \Theta$, we denote the true value of parameter θ by $\theta_0 \in K$. Define the random function

$$\mathbb{Y}_n(\theta; \theta_0) := -\frac{1}{n}\big(Z_n(\theta) - Z_n(\theta_0)\big).$$

Also, let $\theta \mapsto \mathbb{Y}_0(\theta; \theta_0)$ be a random function. We consider the PLAQ representation of $\mathbb{M}_n$:

$$\mathbb{M}_n(w; \theta_0) = \Delta_n(\theta_0)[w] + \frac{1}{2}\Gamma_0(\theta_0)[w, w] + r_n(w; \theta_0) \tag{1.9}$$

for $w \in \{w \in \mathbb{R}^p : \ \theta_0 + w/\sqrt{n} \in \Theta\}$, where $\Delta_n(\theta_0) \in \mathbb{R}^p$, $\Gamma_0(\theta_0) \in \mathbb{R}^p \times \mathbb{R}^p$ and $r_n(w;\theta_0) \in \mathbb{R}$ are random variables[1]. Finally, let $\alpha \in (0,1)$, $U_n(r,\theta_0) := \{w \in \mathbb{R}^p : \ r \le |w| \le n^{(1-\alpha)/2}\}$. We now introduce some conditions.

[A1] $\exists \nu_1 > 0, \ \forall L > 0, \ \exists c_L > 0 : \text{constant}, \ \forall r > 0,$

$$\sup_{\theta_0 \in K} \sup_{n>0} P\Big(\sup_{w \in U_n(r,\theta_0)} \frac{|r_n(w;\theta_0)|}{1+|w|^2} \ge r^{-\nu_1} \Big) \le \frac{c_L}{r^L}.$$

[A2] $\Gamma_0(\theta_0)$ is deterministic and positive-definite uniformly in $\theta_0 \in K$.

[A3] $\exists \chi = \chi(\theta_0) > 0 : \text{ non-random}, \ \exists \nu = \nu(\theta_0) > 0, \ \forall \theta \in \Theta,$

$$\mathbb{Y}_0(\theta;\theta_0) \le -\chi |\theta - \theta_0|^{\nu}.$$

[A4] $\alpha \in (0,1), \ \nu_1 \in (0,1), \ \alpha\nu < \nu_2, \ \beta \in [0,\infty), \ 1 - 2\beta - \nu_2 > 0.$

[A5] $\forall L > 0, \ N_1 := L(1-\nu_1)^{-1}, \ N_2 := L(1-2\beta-\nu_2)^{-1},$

$$\sup_{\theta_0 \in K} \sup_{n>0} E\Big[|\Delta_n(\theta_0)|^{N_1} \Big] < \infty;$$

$$\sup_{\theta_0 \in K} \sup_{n>0} E\Big[\Big(\sup_{\theta \in \Theta} n^{1/2-\beta} \big| \mathbb{Y}_n(\theta;\theta_0) - \mathbb{Y}_0(\theta;\theta_0) \big| \Big)^{N_2} \Big] < \infty.$$

Theorem 1.1 (Yoshida [12], Theorems 1 and 3(a)) *Assume [A1]–[A5]. Then, the estimate* (1.8) *holds uniformly in* $\theta_0 \in K$.

2 Moment convergence

In this section we will deduce the PLDI for the regularized-LSE $\hat{\theta}_n = (\hat{z}_n, \hat{\rho}_n)$. In Section 2.1, we will derive the moment convergence of $\hat{w}_n$. Section 2.2 will discuss the partial PLDI for zero parameter z under different conditions, regarding the non-zero parameter ρ as a nuisance parameter.

2.1 Joint PLDI

In this section we discuss the moment convergence of $\hat{w}_n$ by checking the conditions of Theorem 1.1. In particular, if we have the weak convergence $\hat{w}_n \xrightarrow{\mathcal{L}} \hat{w}_0$ for some random vector $\hat{w}_0$, then the moment convergence (1.7) holds. Let $C_n := n^{-1} \sum_{i=1}^n X_i X_i^\top$.

Theorem 2.1 *Assume that the linear regression model is* (1.2) *and the contrast function is* (1.3). *Suppose the following conditions.*

$$\epsilon_1, \epsilon_2, \ldots \textit{ are i.i.d. with } E[\epsilon_i] = 0 \textit{ and } \forall k > 0, \ E[|\epsilon_i|^k] < \infty; \qquad (2.1)$$

[1] The sign in front of the quadratic term $(1/2)\Gamma_0(\theta_0)[w,w]$ is different from the original PLAQ of Yoshida [12] since we consider minimization of (1.4).

$$\exists \delta > 0,\ \exists C_0 > 0,\ \sup_{n>0} \big(n^{\delta} |C_n - C_0| \big) < \infty; \tag{2.2}$$

$$\sup_{n>0} \sup_{i \le n} |X_i| < \infty; \tag{2.3}$$

$$\exists \beta \in \Big(0, \frac{1}{2}\Big),\ \forall a \in \mathbb{R},\ \sup_{n>0} \frac{\mathfrak{p}_n(a)}{n^{1/2+\beta}} < \infty; \tag{2.4}$$

$$\begin{aligned} &\exists \kappa \in (0,2),\ \forall a \neq 0,\ \exists c_a > 0 \text{ constant},\ \forall b \in \mathbb{R}, \\ &\limsup_{n\to\infty} \Big| \mathfrak{p}_n\Big(a + \frac{b}{\sqrt{n}}\Big) - \mathfrak{p}_n(a) \Big| \le c_a |b|^{\kappa}. \end{aligned} \tag{2.5}$$

Then the PLDI (1.8) *holds. Additionally if we have the weak convergence* $\hat{w}_n \xrightarrow{\mathcal{L}} \hat{w}_0$ *for some random vector* $\hat{w}_0$*, then the moment convergence* (1.7) *holds.*

Proof We will check the conditions of Theorem 1.1 to conclude (1.8). Set $w = (u, v) \in \mathbb{R}^{p_0} \times \mathbb{R}^{p_1}$. We have the statistical random fields

$$\begin{aligned} \mathbb{M}_n(w;\theta_0) &= Z_n\Big(\theta_0 + \frac{w}{\sqrt{n}}\Big) - Z_n(\theta_0) \\ &= \sum_{i=1}^{n} \Big\{ \Big(\epsilon_i - \frac{w^{\top}}{\sqrt{n}} X_i \Big)^2 - \epsilon_i^2 \Big\} + \sum_{k=1}^{p_0} \mathfrak{p}_n\Big(\frac{u_k}{\sqrt{n}} \Big) + \sum_{l=1}^{p_1} \Big\{ \mathfrak{p}_n\Big(\rho_{0l} + \frac{v_l}{\sqrt{n}} \Big) - \mathfrak{p}_n(\rho_{0l}) \Big\} \\ &= -\sum_{i=1}^{n} \frac{2}{\sqrt{n}} \epsilon_i X_i[w] + \frac{1}{2}(2C_0)[w,w] + (C_n - C_0)[w,w] + \sum_{k=1}^{p_0} \mathfrak{p}_n\Big(\frac{u_k}{\sqrt{n}} \Big) \\ &\quad + \sum_{l=1}^{p_1} \Big\{ \mathfrak{p}_n\Big(\rho_{0l} + \frac{v_l}{\sqrt{n}} \Big) - \mathfrak{p}_n(\rho_{0l}) \Big\}. \end{aligned}$$

Since $\hat{w}_n$ is a minimum point of $\mathbb{M}_n(w;\theta_0)$ and $\mathfrak{p}_n$ is a non-negative function, we have

$$\begin{aligned} P(|\hat{w}_n| \ge r) &\le P\Big[\sup_{|w|\ge r} \big\{ -\mathbb{M}_n(w;\theta_0) \big\} \ge -\mathbb{M}_n(0;\theta_0) = 0 \Big] \\ &\le P\Big[\sup_{|w|\ge r} \Big\{ \sum_{i=1}^{n} \frac{2}{\sqrt{n}} \epsilon_i X_i[w] - \frac{1}{2}(2C_0)[w,w] - (C_n - C_0)[w,w] \\ &\qquad - \sum_{l=1}^{p_1} \Big(\mathfrak{p}_n\Big(\rho_{0l} + \frac{v_l}{\sqrt{n}} \Big) - \mathfrak{p}_n(\rho_{0l}) \Big) \Big\} \ge 0 \Big]. \end{aligned}$$

Hence, we will establish the PLDI

$$\begin{aligned} &\sup_{n>0} P\Big[\sup_{|w|\ge r} \Big\{ \sum_{i=1}^{n} \frac{2}{\sqrt{n}} \epsilon_i X_i[w] - \frac{1}{2}(2C_0)[w,w] - (C_n - C_0)[w,w] \\ &\qquad - \sum_{l=1}^{p_1} \Big(\mathfrak{p}_n\Big(\rho_{0l} + \frac{v_l}{\sqrt{n}} \Big) - \mathfrak{p}_n(\rho_{0l}) \Big) \Big\} \ge 0 \Big] \le \frac{c_L}{r^L}, \quad r > 0 \end{aligned} \tag{2.6}$$

for any $L > 0$ to ensure the PLDI (1.8). We have the PLAQ structure with

$$\Delta_n(\theta_0) = \sum_{i=1}^{n} \frac{2}{\sqrt{n}} \epsilon_i X_i; \tag{2.7}$$

$$\Gamma_0(\theta_0) = 2C_0; \tag{2.8}$$

$$r_n(w;\theta_0) = -(C_n - C_0)[w,w] - \sum_{l=1}^{p_1} \Big\{ \mathfrak{p}_n\Big(\rho_{0l} + \frac{v_l}{\sqrt{n}}\Big) - \mathfrak{p}_n(\rho_{0l}) \Big\}. \tag{2.9}$$

According to (2.1)–(2.4), we obtain for any $\theta \in \Theta$

$$\begin{aligned}
\mathbb{Y}_n(\theta;\theta_0) &= -\frac{1}{n}\big(Z_n(\theta) - Z_n(\theta_0)\big) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\Big[\big\{\epsilon_i - (\theta-\theta_0)^\top X_i\big\}^2 - \epsilon_i^2\Big] - \frac{1}{n}\sum_{j=1}^{p}\{\mathfrak{p}_n(\theta_j) - \mathfrak{p}_n(\theta_{0j})\} \\
&= \frac{2}{n}\sum_{i=1}^{n}\epsilon_i X_i[\theta-\theta_0] - C_n[\theta-\theta_0,\theta-\theta_0] - \frac{1}{n}\sum_{j=1}^{p}\{\mathfrak{p}_n(\theta_j) - \mathfrak{p}_n(\theta_{0j})\} \\
&\xrightarrow{P} -C_0[\theta-\theta_0,\theta-\theta_0] =: \mathbb{Y}_0(\theta;\theta_0).
\end{aligned}$$

We get $\mathbb{Y}_0(\theta;\theta_0) \le -\lambda_{\min}(C_0)|\theta-\theta_0|^2$ where $\lambda_{\min}(C_0)$ denotes the minimal eigen-value of the matrix C_0. Apparently [A2] holds from (2.2) and (2.8), and also [A3] holds with $\chi = \lambda_{\min}(C_0)$ and $\nu = 2$. Hence it remains to verify [A1], [A4] and [A5].

First, we will verify [A1]. From (2.9), we have

$$\frac{|r_n(w;\theta_0)|}{1+|w|^2} \le \frac{|w|^2}{1+|w|^2}|C_n - C_0| + \frac{1}{1+|w|^2}\Big|\sum_{l=1}^{p_1}\Big\{\mathfrak{p}_n\Big(\rho_{0l}+\frac{v_l}{\sqrt{n}}\Big) - \mathfrak{p}_n(\rho_{0l})\Big\}\Big|. \tag{2.10}$$

Let us fix β, ν_2, $\alpha \in (0,1)$ and ξ such that $0 \vee (1/2-\delta) \le \beta < 1/2$, $1-2\beta > \nu_2 > 2\alpha$ and $0 < \xi < \big(2\alpha/(1-\alpha)\big) \wedge 1$. Note that these parameters meet $\beta - 1/2 + (1-\alpha)\xi/2 < 0$. Then for the first term of the right-hand side of (2.10), we get from (2.2)

$$\begin{aligned}
&\sup_{w\in U_n(r,\theta_0)}\Big(\frac{|w|^2}{1+|w|^2}|C_n - C_0|\Big) \\
&= n^{1/2-\beta-\delta}\big(n^{\delta}|C_n - C_0|\big)\sup_{w\in U_n(r,\theta_0)}\Big(\frac{|w|^2}{1+|w|^2}n^{\beta-1/2}|w|^{\xi}|w|^{-\xi}\Big) \\
&\lesssim n^{\beta-1/2}n^{(1-\alpha)\xi/2}r^{-\xi} \lesssim r^{-\xi},
\end{aligned} \tag{2.11}$$

where $A_n \lesssim B_n$ means that $\sup_n(A_n/B_n) < \infty$. Next we will estimate the second term of the right-hand side of (2.10). We obtain from (2.5) that there exists a $\kappa \in (0,2)$ such that

$$\frac{1}{1+|w|^2}\Big|\sum_{l=1}^{p_1}\Big\{\mathfrak{p}_n\Big(\rho_{0l}+\frac{v_l}{\sqrt{n}}\Big) - \mathfrak{p}_n(\rho_{0l})\Big\}\Big| \lesssim \frac{|v|^{\kappa}}{1+|w|^2} \lesssim |w|^{\kappa-2}, \quad w \in U_n(r,\theta_0);$$

note that $\sup_{w \in U_n(r,\theta_0)} |v_l|/\sqrt{n} \to 0$. Since we can take $\alpha \in (0,1)$ such that $2-\kappa > \xi$ $\big($note that $0 < \xi < \big(2\alpha/(1-\alpha)\big) \wedge 1\big)$, we get

$$\sup_{w \in U_n(r,\theta_0)} |w|^{\kappa-2} \lesssim r^{-\xi}. \tag{2.12}$$

Fix a $\nu_1 \in (0,\xi)$. Then from (2.10)–(2.12), we have for any $L > 0$

$$\sup_{n>0} P\Big(\sup_{w \in U_n(r,\theta_0)} \frac{|r_n(w;\theta_0)|}{1+|w|^2} \gtrsim r^{-\nu_1} \Big) \lesssim \frac{1}{r^L}.$$

This means that [A1] holds, and [A4] also holds with taking the parameters as above.

Second, we will verify [A5]. From (2.7), we define $\Delta_n(\theta_0) = \sum_{i=1}^n (2/\sqrt{n})\epsilon_i X_i =: \sum_{i=1}^n \chi_{ni}$. Then by using Burkholder's inequality and Jensen's inequality we obtain for $N_1 = L(1-\nu_1)^{-1} \geq 2$

$$\begin{aligned}
\sup_{n>0} E\Big[\big|\Delta_n(\theta_0)\big|^{N_1} \Big] &\leq \sup_{n>0} E\Big[\max_{j \leq n} \Big| \sum_{i=1}^{j} \chi_{ni} \Big|^{N_1} \Big] \\
&\lesssim \sup_{n>0} E\Big[\Big(\sum_{i=1}^{n} \chi_{ni}^2 \Big)^{N_1/2} \Big] \\
&\lesssim \sup_{n>0} E\Big[\frac{1}{n} \sum_{i=1}^{n} \big|\epsilon_i X_i\big|^{2 \cdot N_1/2} \Big] \\
&\lesssim E\big[|\epsilon_1|^{N_1} \big] \cdot \sup_{n>0} \Big(\frac{1}{n} \sum_{i=1}^{n} |X_i|^{N_1} \Big) < \infty.
\end{aligned} \tag{2.13}$$

The last boundedness of (2.13) follows from (2.1) and (2.3). Moreover, we get for any $\theta \in \Theta$

$$\sum_{i=1}^{n} \frac{2}{n} \epsilon_i X_i[\theta-\theta_0] - C_n[\theta-\theta_0, \theta-\theta_0] \xrightarrow{P} -C_0[\theta-\theta_0, \theta-\theta_0];$$

$$\frac{1}{n} \sum_{j=1}^{p} \big\{ \mathfrak{p}_n(\theta_j) - \mathfrak{p}_n(\theta_{0j}) \big\} \xrightarrow{P} 0.$$

Since $(a+b)^{N_2} \lesssim a^{N_2} + b^{N_2}$ for any $a, b \geq 0$ and $N_2 = L(1-2\beta-\nu_2)^{-1} \geq 2$, we have

$$\begin{aligned}
&\sup_{n>0} E\Big[\sup_{\theta \in \Theta} \Big(n^{1/2-\beta} \Big| \sum_{i=1}^{n} \frac{2}{n} \epsilon_i X_i[\theta-\theta_0] - C_n[\theta-\theta_0, \theta-\theta_0] + C_0[\theta-\theta_0, \theta-\theta_0] \Big| \Big)^{N_2} \Big] \\
&\lesssim \sup_{n>0} \Big(n^{-\beta N_2} E\Big[\Big| \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \epsilon_i X_i \Big|^{N_2} \Big] \Big) + \Big\{ \sup_{n>0} \big(n^{1/2-\beta-\delta} n^{\delta} |C_n - C_0| \big) \Big\}^{N_2} < \infty.
\end{aligned} \tag{2.14}$$

Note that the parameter space Θ is a compact set. Further, we obtain

$$\sup_{n>0} \sup_{\theta \in \Theta} \Big[n^{1/2-\beta} \Big| \frac{1}{n} \sum_{j=1}^{p} \big\{ \mathfrak{p}_n(\theta_j) - \mathfrak{p}_n(\theta_{0j}) \big\} \Big| \Big]^{N_2} < \infty \tag{2.15}$$

since we have (2.4). From (2.13)–(2.15), we conclude that [A5] holds. Therefore the proof of (1.8) is complete because we established the PLDI (2.6). The latter claim of the theorem is trivial. □

Remark 2.2 We could deal with random design (X_i). Assume for simplicity that (X_i) and (ϵ_j) are independent. Then in order to conclude (1.8), we need to change (2.2) and (2.3) into (2.16) and (2.17), respectively:

$$\exists\delta > 0,\ \exists C_0 > 0 : \text{constant},\ \forall k > 0,\ \sup_{n>0} E\left[\left|n^{\delta}(C_n - C_0)\right|^k\right] < \infty. \tag{2.16}$$

$$\forall k > 0,\ \sup_{n>0}\sup_{i\le n} E\left[|X_i|^k\right] < \infty. \tag{2.17}$$

The corresponding proofs are entirely analogous to the case of deterministic X. □

Remark 2.3 Although we have additionally imposed (2.2) (or (2.16) when X is random), they are automatically satisfied as soon as we may standardize the covariates X_i beforehand: just use $\widetilde{X}_i := C_n^{-1/2} X_i$ instead of the original X_i, so that

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i \widetilde{X}_i^{\top} \equiv I_p \quad (p \times p\text{-identity matrix}).$$

Then (2.2) and (2.16) hold with $C_0 = I_p$. □

2.2 Partial PLDI derivation under different design condition

In this section, we will show that under different set of conditions it is possible to deduce a PLDI for the random fields only associated with the zero parameter z, regarding the non-zero parameter ρ as a nuisance parameter (hence we derived a uniform-in-ρ PLDI) and by utilizing the special nature of the least-squares term. We do not require any information of asymptotic behavior of $n^{-1}\sum_{i=1}^{n} X_i^{(\rho)} X_i^{(\rho)\top}$, which is the $p_1 \times p_1$ submatrix located in the bottom right corner of C_n, but instead we do a kind of orthogonality between $(X_i^{(z)})$ and $(X_i^{(\rho)})$.

Theorem 2.4 *Assume that the linear regression model is* (1.2) *and the contrast function is* (1.3)*. In addition to* (2.1) *and* (2.3)*, we suppose that*

$$\exists D_0 > 0,\ D_n \to D_0, \tag{2.18}$$

where D_n is the $p_0 \times p_0$ submatrix located in the upper left corner of the matrix C_n. Moreover, we suppose that there exist a positive real sequence (q_n) and a positive function $f(r) \to \infty$ as $r \to \infty$, such that

$$\sup_{n>0}\left|\frac{1}{\sqrt{nq_n}}\sum_{i=1}^{n}\left(X_i^{(z)} \otimes X_i^{(\rho)}\right)\right| < \infty; \tag{2.19}$$

$$\inf_{|u|\ge r}\sum_{k=1}^{p_0}\mathfrak{p}_n\Big(\frac{u_k}{\sqrt{n}}\Big)\ge q_n f(r). \tag{2.20}$$

Then for any $L>0$ there exists a constant $c_L>0$ for which

$$\sup_{n>0}P\big(|\sqrt{n}\hat{z}_n|\ge r\big)\le\frac{c_L}{f(r)^L},\quad r>0. \tag{2.21}$$

Proof We have for $u\in\mathbb{R}^{p_0}$

$$\begin{aligned}
&\mathbb{M}_n(u,\rho;\theta_0)\\
&=Z_n\Big(\frac{u}{\sqrt{n}},\rho\Big)-Z_n(0,\rho)\\
&=\sum_{i=1}^n\Big[\Big\{\epsilon_i-\frac{u^\top}{\sqrt{n}}X_i^{(z)}-(\rho-\rho_0)^\top X_i^{(\rho)}\Big\}^2-\big\{\epsilon_i-(\rho-\rho_0)^\top X_i^{(\rho)}\big\}^2\Big]+\sum_{k=1}^{p_0}\mathfrak{p}_n\Big(\frac{u_k}{\sqrt{n}}\Big)\\
&=-\sum_{i=1}^n\frac{2}{\sqrt{n}}\big\{\epsilon_i-(\rho-\rho_0)^\top X_i^{(\rho)}\big\}X_i^{(z)}[u]+D_n[u,u]+\sum_{k=1}^{p_0}\mathfrak{p}_n\Big(\frac{u_k}{\sqrt{n}}\Big).
\end{aligned}$$

In the present case, we can directly estimate of the tail probability by making use of the special nature of the least-squares term. Let

$$S_n^\rho:=\sum_{i=1}^n\frac{2}{\sqrt{n}}\big\{\epsilon_i-(\rho-\rho_0)^\top X_i^{(\rho)}\big\}X_i^{(z)}.$$

Since $Z_n(z,\rho)\ge Z_n(\hat{z}_n,\hat{\rho}_n)$ for any $(z,\rho)\in\Theta_0\times\Theta_1$, $Z_n(0,\hat{\rho}_n)-Z_n(\hat{z}_n,\hat{\rho}_n)\ge 0$ implies that $\sup_{\rho\in\Theta_1}(1/q_n)\big\{Z_n(0,\rho)-Z_n(\hat{z}_n,\rho)\big\}\ge 0$. Hence we get

$$\begin{aligned}
P\big(|\sqrt{n}\hat{z}_n|\ge r\big)&\le P\Big[\sup_{\rho\in\Theta_1}\sup_{|u|\ge r}\Big\{-\frac{1}{q_n}\mathbb{M}_n(u,\rho;\theta_0)\Big\}\ge 0\Big]\\
&\le P\Big\{\sup_{\rho\in\Theta_1}\sup_{|u|\ge r}\Big(\frac{S_n^\rho}{q_n}[u]-\frac{D_n}{q_n}[u,u]\Big)\ge\frac{1}{q_n}\inf_{|u|\ge r}\sum_{k=1}^{p_0}\mathfrak{p}_n\Big(\frac{u_k}{\sqrt{n}}\Big)\Big\}\\
&\lesssim P\Big\{\sup_{\rho\in\Theta_1}\Big|\frac{1}{\sqrt{q_n}}S_n^\rho\Big|^2\ge\frac{1}{q_n}\inf_{|u|\ge r}\sum_{k=1}^{p_0}\mathfrak{p}_n\Big(\frac{u_k}{\sqrt{n}}\Big)\Big\}
\end{aligned}$$

because we have (2.18) and the quadratic function $S_n^\rho/q_n[u]-D_n/q_n[u,u]$ has the maximum value $(1/4q_n)S_n^{\rho\top}D_n^{-1}S_n^\rho\lesssim|q_n^{-1/2}S_n^\rho|^2$. Therefore, according to (2.19) and (2.20), Markov's inequality and the same argument as (2.13) give us

$$\begin{aligned}
P\big(|\sqrt{n}\hat{z}_n|\ge r\big)&\lesssim P\Big(\sup_{\rho\in\Theta_1}\Big|\frac{1}{\sqrt{q_n}}S_n^\rho\Big|^2\gtrsim f(r)\Big)\\
&\lesssim f(r)^{-L}\Big\{\Big|\sum_{i=1}^n\frac{1}{\sqrt{nq_n}}\big(X_i^{(z)}\otimes X_i^{(\rho)}\big)\Big|^{2L}+q_n^{-L}E\Big[\Big|\sum_{i=1}^n\frac{1}{\sqrt{n}}\epsilon_iX_i^{(z)}\Big|^{2L}\Big]\Big\}\\
&\lesssim f(r)^{-L}.
\end{aligned}$$

Hence we get the PLDI (2.21). □

Remark 2.5 As was mentioned in Remark 2.3, the condition (2.19) is automatic and is not real restrictions if $X_i, \ldots, X_n$ are standardized so that $C_n = I_p$ from the beginning. Then (2.19) is never a real restriction. □

Remark 2.6 We could drive the PLDI for the random fields only associated with the non-zero parameter ρ, regarding the zero parameter z as a nuisance parameter (hence we derived a uniform-in-z PLDI); in this case, we do not impose any condition on the asymptotic behavior of $D_n = n^{-1}\sum_{i=1}^{n} X_i^{(z)} X_i^{(z)\top}$. This can be proved by making use of the argument of Yoshida [12] under the conditions including (2.4), (2.5) and the condition stronger than (2.19):

$$\sup_{n>0}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i^{(z)} \otimes X_i^{(\rho)})\right| < \infty.$$

□

3 Examples

We will give some examples of the regularization term in (1.3) satisfying the conditions (2.4) and (2.5) in Theorem 2.1: sparse-bridge (Radchenko [6]), the smoothly clipped absolute deviation (SCAD; Fan and Li [2]) and the seamless-L_0 regularization (Dicker et al. [1]). From the previous studies, it is known that these regularized estimators $\hat{\theta}_n = (\hat{z}_n, \hat{\rho}_n)$ have the sparse consistency $P(\hat{z}_n = 0) \to 1$, which concludes the sparse estimation, and the asymptotic laws of $\sqrt{n}(\hat{\rho}_n - \rho_0)$ under some appropriate regularity conditions. Also when the number of variables $p = p_n \to \infty$ as $n \to \infty$, the asymptotic behavior of the SCAD and the seamless-L_0 estimators are known, but once again, note that we consider the case that p is fixed.

3.1 Sparse-bridge

In this section we will focus on the sparse-bridge LSE defined the contrast function to be

$$Z_n(\theta) = Z_n(z, \rho) := \sum_{i=1}^{n}(Y_i - z^\top X_i^{(z)} - \rho^\top X_i^{(\rho)})^2 + \lambda_n \sum_{j=1}^{p}|\theta_j|^\gamma, \tag{3.1}$$

where $\lambda_n \geq 0$ denotes the tuning parameter controlling the degree of regularization together with the bridge index $\gamma \in (0, 1)$. This means $\mathfrak{p}_n(\cdot) = \lambda_n|\cdot|^\gamma$. Denote by $\hat{\theta}_n = (\hat{z}_n, \hat{\rho}_n)$ a minimizer of Z_n over a compact parameter space $\Theta = \Theta_0 \times \Theta_1 \subset \mathbb{R}^{p_0} \times \mathbb{R}^{p_1}$. The asymptotic behavior of $\hat{\theta}_n$ is studied by Radchenko [6]. They assumed regularity conditions including that the noises $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. with $E[\epsilon_i] = 0$ and $E[\epsilon_i^2] =: \sigma^2 > 0$, $C_n \to C_0$ for some $C_0 > 0$ and that $n^{-1}\max_{i\leq n}|X_i|^2 \to 0$. Note that these conditions are satisfied with (2.1)–(2.3). Then they proved the following results:

- The sparse consistency of $\hat{z}_n$:

$$P(\hat{z}_n = 0) \to 1 \text{ if } \lambda_n/n^{\gamma/2} \to \infty \text{ and } \lambda_n/n \to 0.$$

- The asymptotic laws of $\hat{\rho}_n$:

(i) $\sqrt{n}(\hat{\rho}_n - \rho_0) \xrightarrow{\mathcal{L}} N_{p_1}(-\lambda_0 B_0^{-1}\Upsilon, \sigma^2 B_0^{-1})$ if $\lambda_n/n^{\gamma/2} \to \infty$ and $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$;

(ii) $n\lambda_n^{-1}(\hat{\rho}_n - \rho_0) \xrightarrow{\mathcal{L}} -B_0^{-1}\Upsilon$ if $\lambda_n/\sqrt{n} \to \infty$ and $\lambda_n/n \to 0$,

where

$$\Upsilon := \frac{\gamma}{2}\{\mathrm{sgn}(\rho_{0,1})|\rho_{0,1}|^{\gamma-1}, \ldots, \mathrm{sgn}(\rho_{0,p_1})|\rho_{0,p_1}|^{\gamma-1}\}$$

and B_0 is the $p_1 \times p_1$ submatrix located in the bottom right corner of the matrix C_0. We are concerned here with the moment convergence of $\hat{w}_n$. With regard to the asymptotic law of the non-zero parameter ρ, we only consider the case (i), where the asymptotic distribution is non-degenerate. The following Corollary 3.1 is derived from Theorem 2.1.

Corollary 3.1 *Assume that the linear regression model is* (1.2) *and the contrast function is* (3.1)*, where* $\gamma \in (0,1)$*,* $\lambda_n/n^{\gamma/2} \to \infty$ *and* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$*. Suppose that* (2.1)*–*(2.3)*. Then the PLDI* (1.8) *holds. In particular, the moment convergence* (1.7) *holds with* $\hat{w}_0 = (0, \hat{v}_0)$*, where* $\mathcal{L}(\hat{v}_0) = N_{p_1}(-\lambda_0 B_0^{-1}\Upsilon, \sigma^2 B_0^{-1})$.

Proof Apparently, we only need to check the conditions (2.4) and (2.5) in Theorem 2.1 for $\mathfrak{p}_n(\cdot) = \lambda_n|\cdot|^\gamma$. (2.4) follows easily since for any $a \in \mathbb{R}$, we have

$$\frac{\mathfrak{p}_n(a)}{\sqrt{n}} = \frac{\lambda_n}{\sqrt{n}}|a|^\gamma \lesssim 1$$

from $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$. We will show (2.5). When n is large enough, we have for any $a \neq 0$ and $b \in \mathbb{R}$

$$\begin{aligned}\left|\mathfrak{p}_n\left(a + \frac{b}{\sqrt{n}}\right) - \mathfrak{p}_n(a)\right| &= \lambda_n\left||a + \frac{b}{\sqrt{n}}|^\gamma - |a|^\gamma\right| \\ &\lesssim \frac{\lambda_n}{\sqrt{n}}|b| \lesssim |b|.\end{aligned}$$

This shows that (2.5) holds for $\kappa = 1$, hence we obtain the PLDI (1.8). The latter claim is trivial since we have $(\sqrt{n}\hat{z}_n, \sqrt{n}(\hat{\rho}_n - \rho_0)) \xrightarrow{\mathcal{L}} (0, \hat{v}_0)$, where $\mathcal{L}(\hat{v}_0) = N_{p_1}(-\lambda_0 B_0^{-1}\Upsilon, \sigma^2 B_0^{-1})$. □

In Section 2.2 we considered the PLDI for the random fields only associated with the zero parameter z. In the following Corollary 3.2, we derive this partial PLDI for the sparse-bridge LSE. It is a direct corollary of Theorem 2.4, so we omit the proof.

Corollary 3.2 *Assume that the linear regression model is* (1.2) *and the contrast function is* (3.1)*, where* $\gamma \in (0,1)$*,* $\lambda_n/n^{\gamma/2} \to \infty$ *and* $\lambda_n/n \to 0$*. Suppose that* (2.1)*,* (2.3)*,* (2.18) *and*

$$\sup_{n>0}\left|\left(\lambda_n/n^{\gamma/2}\right)^{-1/2} n^{-1/2}\sum_{i=1}^{n}\left(X_i^{(z)} \otimes X_i^{(\rho)}\right)\right| < \infty.$$

Then the PLDI (2.21) *holds with* $f(r) = r^\gamma$.

Remark 3.3 Here, we briefly mention the case of the bridge-LSE $\hat{\theta}_n$ defined as the minimal point of the contrast function

$$Z_n(\theta) := \sum_{i=1}^{n}(Y_i - \theta^\top X_i)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|^\gamma, \tag{3.2}$$

where $\lambda_n \geq 0$ and $\gamma > 0$ satisfy that $\lambda_n/n^{(1\wedge\gamma)/2} \to \lambda_0 \geq 0$; then, we do not have the sparse consistency. Note that, different from (3.1), in (3.2) we do not divide the true value of parameter θ_0 into the zero part and the non-zero part: jointly estimate all the components. We assume (2.1)–(2.3). Then Knight and Fu [3] proved the following asymptotic behavior of $\hat{\theta}_n$.

- Consistency:

$$\hat{\theta}_n \xrightarrow{P} \theta_0 \text{ if } \lambda_n/n \to 0.$$

- Asymptotic laws:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \operatorname{argmin}(V_0) \text{ if } \lambda_n/n^{(1\wedge\gamma)/2} \to \lambda_0 \geq 0,$$

  where for $W \sim N_p(0, \sigma^2 C_0)$,

$$V_0(w) := \begin{cases} -2W[w] + C_0[w,w] + \gamma\lambda_0 \sum_{j=1}^{p} w_j \operatorname{sgn}(\theta_{0j})|\theta_{0j}|^{\gamma-1} & (\gamma > 1), \\ -2W[w] + C_0[w,w] & \\ \quad +\lambda_0 \sum_{j=1}^{p} \{w_j \operatorname{sgn}(\theta_{0j}) I(\theta_{0j} \neq 0) + |w_j| I(\theta_{0j} = 0)\} & (\gamma = 1), \\ -2W[w] + C_0[w,w] + \lambda_0 \sum_{j=1}^{p} |w_j|^\gamma I(\theta_{0j} = 0) & (\gamma < 1). \end{cases}$$

Let $\hat{w}_0 = \operatorname{argmin}(V_0)$. We can derive the PLDI for the bridge-LSE by making use of the argument similar to the proof of Theorem 2.1: for any $L > 0$ there exists a constant $c_L > 0$ for which $\sup_{n>0} P(|\hat{w}_n| \geq r) \leq c_L r^{-L}$ $(r > 0)$. In particular, for every continuous $f : \mathbb{R}^p \to \mathbb{R}$ of at most polynomial growth, $E[f(\hat{w}_n)] \to E[f(\hat{w}_0)]$. See Masuda and Shimizu [5, Section 2] for details. □

Remark 3.4 As noted by Knight and Fu [3], the bridge-LSE $\hat{\theta}_n$ is inconsistent when $\lambda_n/n \to \exists\lambda_0 > 0$, and instead tends in probability to

$$\theta_0' := \operatorname*{argmin}_{\theta\in\Theta} \Big\{(\theta - \theta_0)^\top C_0(\theta - \theta_0) + \lambda_0 \sum_{j=1}^{p} |\theta_j|^\gamma\Big\}.$$

Even in this case it is possible to derive the sparse consistency and the associated PLDI for the quasi-zero parameters (whenever exist): specifically, assuming that $\theta_0' = (z_0', \rho_0') = (0, \rho_0') \in \Theta_0' \times \Theta_1' \subset \mathbb{R}^{p_0'} \times \mathbb{R}^{p_1'}$, we could prove the convergence $P(\hat{z}_n' = 0) \to 1$ and the PLDI for $\sqrt{n}\hat{z}_n'$, where $\hat{z}_n'$ denotes the bridge-LSE of z_0', by making use of the same argument as in the proof of Theorem 2.4 and Radchenko [7, Theorem 2]. See Masuda and Shimizu [5, Section 4] for details. □

3.2 SCAD

For simplicity, in this section we assume that the covariates X_i are standardized such that $C_n = I_p$. The SCAD-LSE (Fan and Li [2]) is defined as the minimum point of the contrast function (1.3), where

$$\mathfrak{p}_n(\theta_j) = \begin{cases} n\lambda_n|\theta_j| & (|\theta_j| \le \lambda_n), \\ \dfrac{-n(\theta_j^2 - 2\tau\lambda_n|\theta_j| + \lambda_n^2)}{2(\tau-1)} & (\lambda_n < |\theta_j| \le \tau\lambda_n), \\ \dfrac{n(\tau+1)\lambda_n^2}{2} & (|\theta_j| > \tau\lambda_n). \end{cases}$$

$\tau > 2$ is an additional tuning parameter. Let the minimizer be $\hat{\theta}_n = (\hat{z}_n, \hat{\rho}_n)$ and (2.1) hold. Moreover, Fan and Li [2] assumed that the tuning parameter λ_n satisfied

$$\lambda_n \to 0, \ \sqrt{n}\lambda_n \to \infty. \tag{3.3}$$

Then they proved the sparse consistency and the asymptotic law of ρ_n:

$$\sqrt{n}(\hat{\rho}_n - \rho_0) \xrightarrow{\mathcal{L}} N_{p_1}(0, \mathcal{I}_{p_1}(\rho_0)^{-1}),$$

where $\mathcal{I}_{p_1}(\rho_0) = \mathcal{I}_{p_1}(0, \rho_0)$ denotes the $p_1 \times p_1$ Fisher information matrix knowing $z_0 = 0$.

Let us take $\lambda_n \sim n^{\beta-1/2}$, where β is the same as in the proof of Theorem 2.1. This meets (3.3). Now, we will show (2.4) and (2.5). First we establish (2.4). Obviously, we only need to consider the case $\lambda_n < |\theta_j| \le \tau\lambda_n$. When n is large enough, we have $n\theta_j^2/n^{\beta+1/2} \lesssim n\lambda_n^2/n^{\beta+1/2} \sim n^{1+2\beta-1-\beta-1/2} = n^{\beta-1/2} \lesssim 1$, hence (2.4) holds. In order to ensure (2.5), we use

$$\mathfrak{p}_n'(\theta_j) = \lambda_n n\Big\{I(\theta_j \le \lambda_n) + \frac{(\tau\lambda_n - \theta_j)_+}{(\tau-1)\lambda_n} I(\theta_j > \lambda_n)\Big\}, \quad \theta_j > 0.$$

When n is large enough, for any $a > 0$ and $b \in \mathbb{R}$

$$\begin{aligned} \Big|\mathfrak{p}_n\Big(a + \frac{b}{\sqrt{n}}\Big) - \mathfrak{p}_n(a)\Big| &\le \frac{|b|}{\sqrt{n}} \int_0^1 \Big|\mathfrak{p}_n'\Big(a + \frac{b}{\sqrt{n}}t\Big)\Big| dt \\ &\sim \lambda_n\sqrt{n}|b| \int_0^1 I\Big(a + \frac{b}{\sqrt{n}}t \le \lambda_n\Big) dt \\ &\quad + \sqrt{n}|b| \int_0^1 \frac{[\tau\lambda_n - \{a + (b/\sqrt{n})t\}]_+}{\tau - 1} I\Big(a + \frac{b}{\sqrt{n}}t > \lambda_n\Big) dt \\ &\lesssim |b|. \end{aligned}$$

Similarly, we get the same estimate for $a < 0$. As the results, it is possible to take λ_n ensuring (1.7), where $\hat{w}_0 = (0, \hat{v}_0)$ and $\mathcal{L}(\hat{v}_0) = N_{p_1}(0, \mathcal{I}_{p_1}(\rho_0)^{-1})$.

3.3 seamless-L_0

The seamless-L_0 regularization (Dicker et al. [1]), which approximates the (technically unpleasant due to its discontinuity at the origin) L_0-loss, is given by

$$Z_n(\theta) = Z_n(z, \rho) := \sum_{i=1}^n (Y_i - z^\top X_i^{(z)} - \rho^\top X_i^{(\rho)})^2 + \frac{2n\lambda_n}{\log 2} \sum_{j=1}^p \log\Big(\frac{|\theta_j|}{|\theta_j| + \tau_n} + 1\Big),$$

where $\tau_n > 0$ is an additional tuning parameter. Let the minimizer be $\hat{\theta}_n = (\hat{z}_n, \hat{\rho}_n)$ and (2.1)–(2.3) hold. Moreover, Dicker et al. [1] assumed that the tuning parameters satisfied

$$\lambda_n = O(1),\ \lambda_n \sqrt{n} \to \infty,\ \tau_n = O(n^{-3/2}). \tag{3.4}$$

Then they proved the sparse consistency and the asymptotic law of $\hat{\rho}_n$:

$$\sqrt{n}(\hat{\rho}_n - \rho_0) \xrightarrow{\mathcal{L}} N_{p_1}(0, \sigma^2 B_0^{-1}),$$

where B_0 is the same as in Section 3.1.

Let us take $\lambda_n \sim n^{\beta-1/2}$ and $\tau_n \sim n^{-3/2}$, where β is the same as in the proof of Theorem 2.1. This meets (3.4). Now, we will show (2.4) and (2.5) for $\mathfrak{p}_n(\cdot) = (2n\lambda_n/\log 2)\log\{|\cdot|/(|\cdot|+\tau_n)+1\}$. (2.4) follows easily since $\mathfrak{p}_n/n^{1/2+\beta} \lesssim n^{1+\beta-1/2-1/2-\beta} = 1$. In order to ensure (2.5), we make use of the equation

$$|\log(1+x) - \log(1+x')| = \Big| \int_0^1 \frac{ds}{1+x'+(x-x')s}(x-x') \Big|$$

where $x, x' > 0$. When n is large enough, for any $a > 0$ and $b \in \mathbb{R}$

$$\begin{aligned}
\Big|\mathfrak{p}_n\Big(a + \frac{b}{\sqrt{n}}\Big) - \mathfrak{p}_n(a)\Big| &= \frac{2n\lambda_n}{\log 2}\Big|\log\Big(\frac{|a+b/\sqrt{n}|}{|a+b/\sqrt{n}|+\tau_n}+1\Big) - \log\Big(\frac{|a|}{|a|+\tau_n}+1\Big)\Big| \\
&\lesssim n\lambda_n \Big|\frac{a+\delta}{a+\delta+\tau_n} - \frac{a}{a+\tau_n}\Big| \quad (\delta := b/\sqrt{n}) \\
&= n\lambda_n \frac{|(a+\delta)(a+\tau_n) - a(a+\delta+\tau_n)|}{(a+\delta+\tau_n)(a+\tau_n)} \\
&= n\lambda_n \frac{\tau_n|\delta|}{(a+\delta+\tau_n)(a+\tau_n)} \\
&\sim n^{\beta-3/2}|b| \lesssim |b|.
\end{aligned}$$

Similarly, we get the same estimate for $a < 0$. As the results, it is possible to take the tuning parameters ensuring (1.7), where $\hat{w}_0 = (0, \hat{v}_0)$ and $\mathcal{L}(\hat{v}_0) = N_{p_1}(0, \sigma^2 B_0^{-1})$.

Acknowledgements The author is grateful to Professor H. Masuda for his valuable comments.

References

[1] Dicker, L., Huang, B., Lin, X. (2012). Variable selection and estimation with the seamless-L_0 penalty. *Statistica Sinica*, 23(2), 929-962.

[2] Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348-1360.

[3] Knight, K., Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 1356-1378.

[4] Masuda, H. (2013). Convergence of Gaussian quasi-likelihood random fields for ergodic Lévy driven SDE observed at high frequency. *The Annals of Statistics*, 41(3), 1593-1641.

[5] Masuda, H., Shimizu, Y. (2014). Moment convergence in regularized estimations. arXiv:1406.6751v2

[6] Radchenko, P. (2005). Reweighting the lasso. *2005 Proceedings of the American Statistical Association [CD-ROM]*, http://www-rcf.usc.edu/~radchenk/Lasso.pdf

[7] Radchenko, P. (2008). Mixed-rates asymptotics. *The Annals of Statistics*, 36(1), 287-309.

[8] Sakamoto, Y., Yoshida, N. (2004). Asymptotic expansion formulas for functionals of ϵ-Markov processes with a mixing property. *Annals of the Institute of Statistical Mathematics*, 56(3), 545-597.

[9] Uchida, M., Yoshida, N. (2001). Information criteria in model selection for mixing processes. *Statistical Inference for Stochastic Processes*, 4(1), 73-98.

[10] Uchida, M., Yoshida, N. (2006). Asymptotic expansion and information criteria. *SUT Journal of Mathematics*, 42(1), 31-58.

[11] van der Vaart, A. W. (1998). Asymptotic Statistics. *Cambridge University Press.*

[12] Yoshida, N. (2011). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Annals of the Institute of Statistical Mathematics*, 63(3), 431-479.

List of MI Preprint Series, Kyushu University

The Global COE Program
Math-for-Industry Education & Research Hub

MI

MI2008-1 Takahiro ITO, Shuichi INOKUCHI & Yoshihiro MIZOGUCHI
Abstract collision systems simulated by cellular automata

MI2008-2 Eiji ONODERA
The intial value problem for a third-order dispersive flow into compact almost Hermitian manifolds

MI2008-3 Hiroaki KIDO
On isosceles sets in the 4-dimensional Euclidean space

MI2008-4 Hirofumi NOTSU
Numerical computations of cavity flow problems by a pressure stabilized characteristic-curve finite element scheme

MI2008-5 Yoshiyasu OZEKI
Torsion points of abelian varieties with values in nfinite extensions over a p-adic field

MI2008-6 Yoshiyuki TOMIYAMA
Lifting Galois representations over arbitrary number fields

MI2008-7 Takehiro HIROTSU & Setsuo TANIGUCHI
The random walk model revisited

MI2008-8 Silvia GANDY, Masaaki KANNO, Hirokazu ANAI & Kazuhiro YOKOYAMA
Optimizing a particular real root of a polynomial by a special cylindrical algebraic decomposition

MI2008-9 Kazufumi KIMOTO, Sho MATSUMOTO & Masato WAKAYAMA
Alpha-determinant cyclic modules and Jacobi polynomials

MI2008-10 Sangyeol LEE & Hiroki MASUDA
Jarque-Bera Normality Test for the Driving Lévy Process of a Discretely Observed Univariate SDE

MI2008-11 Hiroyuki CHIHARA & Eiji ONODERA
A third order dispersive flow for closed curves into almost Hermitian manifolds

MI2008-12 Takehiko KINOSHITA, Kouji HASHIMOTO and Mitsuhiro T. NAKAO
On the L^2 a priori error estimates to the finite element solution of elliptic problems with singular adjoint operator

MI2008-13 Jacques FARAUT and Masato WAKAYAMA
Hermitian symmetric spaces of tube type and multivariate Meixner-Pollaczek polynomials

MI2008-14 Takashi NAKAMURA
Riemann zeta-values, Euler polynomials and the best constant of Sobolev inequality

MI2008-15 Takashi NAKAMURA
Some topics related to Hurwitz-Lerch zeta functions

MI2009-1 Yasuhide FUKUMOTO
Global time evolution of viscous vortex rings

MI2009-2 Hidetoshi MATSUI & Sadanori KONISHI
Regularized functional regression modeling for functional response and predictors

MI2009-3 Hidetoshi MATSUI & Sadanori KONISHI
Variable selection for functional regression model via the L_1 regularization

MI2009-4 Shuichi KAWANO & Sadanori KONISHI
Nonlinear logistic discrimination via regularized Gaussian basis expansions

MI2009-5 Toshiro HIRANOUCHI & Yuichiro TAGUCHII
Flat modules and Groebner bases over truncated discrete valuation rings

MI2009-6 Kenji KAJIWARA & Yasuhiro OHTA
Bilinearization and Casorati determinant solutions to non-autonomous 1+1 dimensional discrete soliton equations

MI2009-7 Yoshiyuki KAGEI
Asymptotic behavior of solutions of the compressible Navier-Stokes equation around the plane Couette flow

MI2009-8 Shohei TATEISHI, Hidetoshi MATSUI & Sadanori KONISHI
Nonlinear regression modeling via the lasso-type regularization

MI2009-9 Takeshi TAKAISHI & Masato KIMURA
Phase field model for mode III crack growth in two dimensional elasticity

MI2009-10 Shingo SAITO
Generalisation of Mack's formula for claims reserving with arbitrary exponents for the variance assumption

MI2009-11 Kenji KAJIWARA, Masanobu KANEKO, Atsushi NOBE & Teruhisa TSUDA
Ultradiscretization of a solvable two-dimensional chaotic map associated with the Hesse cubic curve

MI2009-12 Tetsu MASUDA
Hypergeometric τ-functions of the q-Painlevé system of type $E_8^{(1)}$

MI2009-13 Hidenao IWANE, Hitoshi YANAMI, Hirokazu ANAI & Kazuhiro YOKOYAMA
A Practical Implementation of a Symbolic-Numeric Cylindrical Algebraic Decomposition for Quantifier Elimination

MI2009-14 Yasunori MAEKAWA
On Gaussian decay estimates of solutions to some linear elliptic equations and its applications

MI2009-15 Yuya ISHIHARA & Yoshiyuki KAGEI
Large time behavior of the semigroup on L^p spaces associated with the linearized compressible Navier-Stokes equation in a cylindrical domain

MI2009-16 Chikashi ARITA, Atsuo KUNIBA, Kazumitsu SAKAI & Tsuyoshi SAWABE
Spectrum in multi-species asymmetric simple exclusion process on a ring

MI2009-17 Masato WAKAYAMA & Keitaro YAMAMOTO
Non-linear algebraic differential equations satisfied by certain family of elliptic functions

MI2009-18 Me Me NAING & Yasuhide FUKUMOTO
Local Instability of an Elliptical Flow Subjected to a Coriolis Force

MI2009-19 Mitsunori KAYANO & Sadanori KONISHI
Sparse functional principal component analysis via regularized basis expansions and its application

MI2009-20 Shuichi KAWANO & Sadanori KONISHI
Semi-supervised logistic discrimination via regularized Gaussian basis expansions

MI2009-21 Hiroshi YOSHIDA, Yoshihiro MIWA & Masanobu KANEKO
Elliptic curves and Fibonacci numbers arising from Lindenmayer system with symbolic computations

MI2009-22 Eiji ONODERA
A remark on the global existence of a third order dispersive flow into locally Hermitian symmetric spaces

MI2009-23 Stjepan LUGOMER & Yasuhide FUKUMOTO
Generation of ribbons, helicoids and complex scherk surface in laser-matter Interactions

MI2009-24 Yu KAWAKAMI
Recent progress in value distribution of the hyperbolic Gauss map

MI2009-25 Takehiko KINOSHITA & Mitsuhiro T. NAKAO
On very accurate enclosure of the optimal constant in the a priori error estimates for H_0^2-projection

MI2009-26 Manabu YOSHIDA
Ramification of local fields and Fontaine's property (Pm)

MI2009-27 Yu KAWAKAMI
Value distribution of the hyperbolic Gauss maps for flat fronts in hyperbolic three-space

MI2009-28 Masahisa TABATA
Numerical simulation of fluid movement in an hourglass by an energy-stable finite element scheme

MI2009-29 Yoshiyuki KAGEI & Yasunori MAEKAWA
Asymptotic behaviors of solutions to evolution equations in the presence of translation and scaling invariance

MI2011-5 Hiroshi INOUE
Implications in Compressed Sensing and the Restricted Isometry Property

MI2011-6 Daeju KIM & Sadanori KONISHI
Predictive information criterion for nonlinear regression model based on basis expansion methods

MI2011-7 Shohei TATEISHI, Chiaki KINJYO & Sadanori KONISHI
Group variable selection via relevance vector machine

MI2011-8 Jan BREZINA & Yoshiyuki KAGEI
Decay properties of solutions to the linearized compressible Navier-Stokes equation around time-periodic parallel flow
Group variable selection via relevance vector machine

MI2011-9 Chikashi ARITA, Arvind AYYER, Kirone MALLICK & Sylvain PROLHAC
Recursive structures in the multispecies TASEP

MI2011-10 Kazunori YASUTAKE
On projective space bundle with nef normalized tautological line bundle

MI2011-11 Hisashi ANDO, Mike HAY, Kenji KAJIWARA & Tetsu MASUDA
An explicit formula for the discrete power function associated with circle patterns of Schramm type

MI2011-12 Yoshiyuki KAGEI
Asymptotic behavior of solutions to the compressible Navier-Stokes equation around a parallel flow

MI2011-13 Vladimír CHALUPECKÝ & Adrian MUNTEAN
Semi-discrete finite difference multiscale scheme for a concrete corrosion model: approximation estimates and convergence

MI2011-14 Jun-ichi INOGUCHI, Kenji KAJIWARA, Nozomu MATSUURA & Yasuhiro OHTA
Explicit solutions to the semi-discrete modified KdV equation and motion of discrete plane curves

MI2011-15 Hiroshi INOUE
A generalization of restricted isometry property and applications to compressed sensing

MI2011-16 Yu KAWAKAMI
A ramification theorem for the ratio of canonical forms of flat surfaces in hyperbolic three-space

MI2011-17 Naoyuki KAMIYAMA
Matroid intersection with priority constraints

MI2012-1 Kazufumi KIMOTO & Masato WAKAYAMA
Spectrum of non-commutative harmonic oscillators and residual modular forms

MI2012-2 Hiroki MASUDA
Mighty convergence of the Gaussian quasi-likelihood random fields for ergodic Levy driven SDE observed at high frequency

MI2012-3 Hiroshi INOUE
A Weak RIP of theory of compressed sensing and LASSO

MI2012-4 Yasuhide FUKUMOTO & Youich MIE
Hamiltonian bifurcation theory for a rotating flow subject to elliptic straining field

MI2012-5 Yu KAWAKAMI
On the maximal number of exceptional values of Gauss maps for various classes of surfaces

MI2012-6 Marcio GAMEIRO, Yasuaki HIRAOKA, Shunsuke IZUMI, Miroslav KRAMAR, Konstantin MISCHAIKOW & Vidit NANDA
Topological Measurement of Protein Compressibility via Persistence Diagrams

MI2012-7 Nobutaka NAKAZONO & Seiji NISHIOKA
Solutions to a q-analog of Painlevé III equation of type $D_7^{(1)}$

MI2012-8 Naoyuki KAMIYAMA
A new approach to the Pareto stable matching problem

MI2012-9 Jan BREZINA & Yoshiyuki KAGEI
Spectral properties of the linearized compressible Navier-Stokes equation around time-periodic parallel flow

MI2012-10 Jan BREZINA
Asymptotic behavior of solutions to the compressible Navier-Stokes equation around a time-periodic parallel flow

MI2012-11 Daeju KIM, Shuichi KAWANO & Yoshiyuki NINOMIYA
Adaptive basis expansion via the extended fused lasso

MI2012-12 Masato WAKAYAMA
On simplicity of the lowest eigenvalue of non-commutative harmonic oscillators

MI2012-13 Masatoshi OKITA
On the convergence rates for the compressible
Navier- Stokes equations with potential force

MI2013-1 Abuduwaili PAERHATI & Yasuhide FUKUMOTO
A Counter-example to Thomson-Tait-Chetayev's Theorem

MI2013-2 Yasuhide FUKUMOTO & Hirofumi SAKUMA
A unified view of topological invariants of barotropic and baroclinic fluids and their application to formal stability analysis of three-dimensional ideal gas flows

MI2013-3 Hiroki MASUDA
Asymptotics for functionals of self-normalized residuals of discretely observed stochastic processes

MI2013-4 Naoyuki KAMIYAMA
On Counting Output Patterns of Logic Circuits

MI2013-5 Hiroshi INOUE
RIPless Theory for Compressed Sensing

MI2013-6 Hiroshi INOUE
Improved bounds on Restricted isometry for compressed sensing

MI2013-7 Hidetoshi MATSUI
Variable and boundary selection for functional data via multiclass logistic regression modeling

MI2013-8 Hidetoshi MATSUI
Variable selection for varying coefficient models with the sparse regularization

MI2013-9 Naoyuki KAMIYAMA
Packing Arborescences in Acyclic Temporal Networks

MI2013-10 Masato WAKAYAMA
Equivalence between the eigenvalue problem of non-commutative harmonic oscillators and existence of holomorphic solutions of Heun's differential equations, eigenstates degeneration, and Rabi's model

MI2013-11 Masatoshi OKITA
Optimal decay rate for strong solutions in critical spaces to the compressible Navier-Stokes equations

MI2013-12 Shuichi KAWANO, Ibuki HOSHINA, Kazuki MATSUDA & Sadanori KONISHI
Predictive model selection criteria for Bayesian lasso

MI2013-13 Hayato CHIBA
The First Painleve Equation on the Weighted Projective Space

MI2013-14 Hidetoshi MATSUI
Variable selection for functional linear models with functional predictors and a functional response

MI2013-15 Naoyuki KAMIYAMA
The Fault-Tolerant Facility Location Problem with Submodular Penalties

MI2013-16 Hidetoshi MATSUI
Selection of classification boundaries using the logistic regression

MI2014-1 Naoyuki KAMIYAMA
Popular Matchings under Matroid Constraints

MI2014-2 Yasuhide FUKUMOTO & Youichi MIE
Lagrangian approach to weakly nonlinear interaction of Kelvin waves and a symmetry-breaking bifurcation of a rotating flow

MI2014-3 Reika AOYAMA
Decay estimates on solutions of the linearized compressible Navier-Stokes equation around a Parallel flow in a cylindrical domain

MI2014-4 Naoyuki KAMIYAMA
The Popular Condensation Problem under Matroid Constraints