Ann Inst Stat Math (2017) 69:1141–1154
DOI 10.1007/s10463-016-0577-6

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

Yusuke Shimizu[1]

Received: 4 Jan 2016 / Revised: 4 July 2016 / Published online: 9 August 2016

Abstract In this paper, we study the uniform tail-probability estimates of a regularized least-squares estimator for the linear regression model. We make use of the polynomial type large deviation inequality for the associated statistical random fields, which may not be locally asymptotically quadratic. Our results enable us to verify various arguments requiring convergence of moments of estimator-dependent statistics, such as the mean squared prediction error and the bias correction for AIC-type information criterion.

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

1 Introduction

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

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

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 $(\epsilon_i)_{i=1}^n$ 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, $\{(X_i, Y_i)\}_{i=1}^n$ are standardized from the beginning such a way that

✉ Yusuke Shimizu
y-shimizu@math.kyushu-u.ac.jp

[1] Graduate school of Mathematics, Kyushu University, 744 Motooka Nishi-ku, Fukuoka 819-0395, Japan

$$\sum_{i=1}^{n} Y_i = 0, \quad \sum_{i=1}^{n} X_{i,j} = 0 \text{ and } \frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2 = 1, \quad j \in \{1, \ldots, p\}.$$

For brevity, we omit the dependence of X_i and Y_i on n from the notation so long as there is no confusion (we use the notation X_{in} when we emphasize the dependence of X_i on n). In this paper, we deal with the situation

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

where $z_{0,k} = 0$ and $\rho_{0,l} \neq 0$ for any $k \in \{1, \ldots, p_0\}$ and $l \in \{1, \ldots, p_1\}$; 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) to

$$Y_i = z_0^\top X_i^{(z)} + \rho_0^\top X_i^{(\rho)} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $X_i^{(z)} := (X_{i,1}, \ldots, X_{i,p_0})^\top$ and $X_i^{(\rho)} := (X_{i,p_0+1}, \ldots, X_{i,p_0+p_1})^\top$, 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

$$Z_n(\theta) = Z_n(z, \rho) := \sum_{i=1}^{n} (Y_i - z^\top X_i^{(z)} - \rho^\top X_i^{(\rho)})^2 + \sum_{j=1}^{p} \mathfrak{p}_n(\theta_j) \tag{3}$$

over Θ, where $\mathfrak{p}_n(\cdot)$ is a non-random and non-negative function such that $\mathfrak{p}_n(0) = 0$. Further conditions on $\mathfrak{p}_n$ will be imposed later on. 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 $n \to \infty$, while $\sqrt{n}(\hat{\rho}_n - \rho_0)$ has a non-trivial asymptotic law. The sparse consistency implies that $R_n \hat{z}_n = o_p(1)$ for arbitrary $R_n \to \infty$, for example, sparse-bridge (Radchenko 2005), the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) and the Seamless-L_0 regularization (Dicker et al 2012). In Sect. 3, we will refer some asymptotic behaviors of these regularized estimators.

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 a true parameter value $\theta_0 \in \Theta$ by a minimum point $\hat{\theta}_n$ of an appropriate continuous contrast function $Z_n : \Theta \to \mathbb{R}$. To assess the asymptotic performance of $\hat{\theta}_n$ quantitatively, when $\sqrt{n}$-consistency is concerned, we look at the statistical random fields

$$\mathbb{M}_n(w; \theta_0) := Z_n\left(\theta_0 + \frac{w}{\sqrt{n}}\right) - Z_n(\theta_0), \tag{4}$$

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 (1998). 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 (2011), which estimates the tail of $\mathcal{L}(\hat{w}_n)$ in such a way that

$$\sup_{r>0} \sup_{n>0} r^L P(|\hat{w}_n| \geq r) < \infty \tag{5}$$

for a given $L > 0$, plays an important role. Assume that there exists a random variable $\hat{w}_0$ such that $\hat{w}_n \xrightarrow{\mathcal{L}} \hat{w}_0$. Then, the moment convergence

$$E[|\hat{w}_n|^q] \to E[|\hat{w}_0|^q], \quad q > 0, \tag{6}$$

holds if there exists a $q' > q$ such that $\sup_{n>0} E[|\hat{w}_n|^{q'}] < \infty$. Suppose that the PLDI (5) holds for some $L > q'$. Then, we obtain

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

As the results, we get the moment convergence (6) if we ensure the PLDI (5) for some $L > q$. Hence, the main purpose of this paper is to derive the PLDI (5) with $\hat{w}_n := (\sqrt{n}\hat{z}_n, \sqrt{n}(\hat{\rho}_n - \rho_0))$.

We should mention the importance of convergence of moments: asymptotic behavior of expected values of statistics depending on estimators. It especially serves as a critical tool when, for example, analyzing the mean squared prediction error and the bias correction for information criteria; see Chen and Ing (2011), Afendras and Markatou (2015a, b), Findley and Wei (2002), Uchida and Yoshida (2001, 2006), Sakamoto and Yoshida (2004), as well as Yoshida (2011). Let us consider a typical scenario. If (6) holds with $q = 2$ and $\hat{w}_0 \sim N_p(0, V)$ where V is a $p \times p$-diagonal matrix, the mean squared error of $\hat{\theta}_n$ can be expressed as:

$$E\left[|\hat{\theta}_n - \theta_0|^2\right] = \frac{\mathrm{tr}(V)}{n} + o_p\left(\frac{1}{n}\right),$$

from which the mean squared prediction error can be established with its theoretical justification. Also, the moment convergence provides benefits to AIC type information criteria, which is widely used as a simple and practical estimate of the best model, and is derived from the bias correction procedures. From the point of view of regularized estimation, AIC is used to select tuning parameters contained in regularization terms. Recently, Umezu et al (2015) proposed the bias-corrected AIC for non-concave regularized likelihood estimator of generalized linear model by verifying its moment convergence, and concluded that the proposed AIC performs well through simulation studies. In particular, they studied asymptotic behaviors of the estimator including $E[|\sqrt{n}\hat{z}_n|^2] \to 0$ which cannot be deduced from the sparse consistency, and is needed

to derive AIC for sparse-type estimations. Convergence of moments for regularized estimators gives us the validity of AIC for selecting tuning parameters.

This paper is organized as follows. In Sect. 2, we will derive the PLDI (5) for the regularized-LSE of the linear regression model (2). We will give some examples of the regularization term in the contrast function (3) in Sect. 3.

For convenience of reference, we end this section with stating Theorem 1 and Theorem 3(a) of Yoshida (2011), which will play an essential role in our study. We need to introduce some notation. For any fixed $\theta_0 \in \Theta$, we define a 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 $\mathbb{M}_n$:

$$\mathbb{M}_n(w;\theta_0) = \Delta_n(\theta_0)[w] + \frac{1}{2}\Gamma_0(\theta_0)[w,w] + r_n(w;\theta_0) \tag{7}$$

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) $\exists \nu_1 > 0,\ \forall L > 0,\ \exists c_L > 0$: constant, $\forall r > 0$,

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

(A2) $\Gamma_0(\theta_0)$ is deterministic and positive-definite.
(A3) $\exists \chi = \chi(\theta_0) > 0$: 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) $\forall L > 0,\ N_1 := L(1-\nu_1)^{-1},\ N_2 := L(1-2\beta-\nu_2)^{-1}$,

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

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

Theorem 1 [Yoshida (2011), Theorems 1 and 3(a)] *Assume* [A1]–[A5]. *Then, the estimate* (5) *holds.*[2]

[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 (2011) since we consider minimization of (4).

[2] The uniform (w.r.t. θ_0) evaluation is not in our scope here.

2 Moment convergence

In this section, we discuss the moment convergence of $\hat{w}_n$ by checking the conditions of Theorem 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 (6) holds. Let $C_n := n^{-1}\sum_{i=1}^n X_i X_i^\top$.

Theorem 2 *Assume that the linear regression model is* (2) *and the contrast function is* (3)*. Suppose the following conditions:*

$$\epsilon_1, \epsilon_2, \dots \text{ are i.i.d. with } E[\epsilon_i] = 0 \quad \text{and} \quad \forall k > 0,\ E[|\epsilon_i|^k] < \infty; \tag{8}$$

$$\exists \delta > 0,\ \exists C_0 > 0,\ \sup_{n>0}\left(n^\delta |C_n - C_0|\right) < \infty; \tag{9}$$

$$\sup_{n>0}\sup_{i\le n}|X_{in}| < \infty; \tag{10}$$

$$\exists \beta \in \left(0, \frac{1}{2}\right),\ \forall K \subset \mathbb{R} : \text{compact},\ \sup_{n>0}\sup_{a\in K}\frac{\mathfrak{p}_n(a)}{n^{1/2+\beta}} < \infty; \tag{11}$$

$$\begin{aligned}&\exists \kappa \in (0,2),\ \forall a \neq 0,\ \exists c_a > 0 : \text{constant},\ \forall b \in \mathbb{R},\\ &\limsup_{n\to\infty}\left|\mathfrak{p}_n\left(a + \frac{b}{\sqrt{n}}\right) - \mathfrak{p}_n(a)\right| \le c_a|b|^\kappa.\end{aligned} \tag{12}$$

Then, the PLDI (5) *holds with* $\hat{w}_n = (\sqrt{n}\hat{z}_n, \sqrt{n}(\hat{\rho}_n - \rho_0))$*. 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* (6) *holds.*

Proof We will check the conditions of Theorem 1 to conclude (5). Set $w = (u, v) \in \mathbb{R}^{p_0} \times \mathbb{R}^{p_1}$. We have the statistical random fields

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

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

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

Hence, we will establish the PLDI

$$\begin{aligned} &\sup_{r>0}\sup_{n>0} r^L P\Big[\sup_{|w|\geq r}\Big\{\sum_{i=1}^{n}\frac{2}{\sqrt{n}}\epsilon_i X_i[w] - \frac{1}{2}(2C_0)[w,w] - (C_n - C_0)[w,w] \\ &\quad - \sum_{l=1}^{p_1}\Big(\mathfrak{p}_n\Big(\rho_{0l} + \frac{v_l}{\sqrt{n}}\Big) - \mathfrak{p}_n(\rho_{0l})\Big)\Big\} \geq 0\Big] < \infty \end{aligned} \tag{13}$$

for any $L > 0$ to ensure the PLDI (5). We have the *PLAQ-like* expression which corresponds to (7) with

$$\Delta_n(\theta_0) = \sum_{i=1}^{n}\frac{2}{\sqrt{n}}\epsilon_i X_i; \tag{14}$$

$$\Gamma_0(\theta_0) = 2C_0; \tag{15}$$

$$r_n(w;\theta_0) = -(C_n - C_0)[w,w] - \sum_{l=1}^{p_1}\Big\{\mathfrak{p}_n\Big(\rho_{0l} + \frac{v_l}{\sqrt{n}}\Big) - \mathfrak{p}_n(\rho_{0l})\Big\}. \tag{16}$$

According to (8)–(11), we obtain for any $\theta \in \Theta$

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

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 eigenvalue of the matrix C_0. Apparently, [A2] holds from (9) and (15), 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 (16), we have

$$\frac{|r_n(w;\theta_0)|}{1+|w|^2} \le \frac{|w|^2}{1+|w|^2}|C_n - C_0| + \frac{1}{1+|w|^2}\Big|\sum_{l=1}^{p_1}\Big\{\mathfrak{p}_n\Big(\rho_{0l}+\frac{v_l}{\sqrt{n}}\Big) - \mathfrak{p}_n(\rho_{0l})\Big\}\Big|. \tag{17}$$

Let us fix $\beta, \nu_2, \alpha \in (0,1)$ 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 (17), we get from (9)

$$\begin{aligned}
&\sup_{w\in U_n(r,\theta_0)}\Big(\frac{|w|^2}{1+|w|^2}|C_n - C_0|\Big) \\
&= n^{1/2-\beta-\delta}\big(n^{\delta}|C_n - C_0|\big)\sup_{w\in U_n(r,\theta_0)}\Big(\frac{|w|^2}{1+|w|^2}n^{\beta-1/2}|w|^{\xi}|w|^{-\xi}\Big) \\
&\lesssim n^{\beta-1/2}n^{(1-\alpha)\xi/2}r^{-\xi} \lesssim r^{-\xi},
\end{aligned} \tag{18}$$

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 (17). We obtain from (12) that there exists a $\kappa \in (0,2)$ such that

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

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 $2-\kappa > \xi$ (note that $0 < \xi < \big(2\alpha/(1-\alpha)\big) \wedge 1$), we get

$$\sup_{w\in U_n(r,\theta_0)}|w|^{\kappa-2} \lesssim r^{-\xi}. \tag{19}$$

Fix a $\nu_1 \in (0,\xi)$. Then from (17)–(19), we have for any $L > 0$

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

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

Second, we will verify [A5]. From (14), we define $\Delta_n(\theta_0) = \sum_{i=1}^n (2/\sqrt{n})\epsilon_i X_i =: \sum_{i=1}^n \chi_{ni}$. Then, using Burkholder's inequality and Jensen's inequality, we obtain for $N_1 = L(1-\nu_1)^{-1} \ge 2$

$$\begin{aligned}\sup_{n>0} E\Big[\big|\Delta_n(\theta_0)\big|^{N_1}\Big] &\le \sup_{n>0} E\Big[\max_{j\le n}\Big|\sum_{i=1}^{j}\chi_{ni}\Big|^{N_1}\Big]\\ &\lesssim \sup_{n>0} E\Big[\Big(\sum_{i=1}^{n}\chi_{ni}^2\Big)^{N_1/2}\Big]\\ &\lesssim \sup_{n>0} E\Big[\frac{1}{n}\sum_{i=1}^{n}\big|\epsilon_i X_i\big|^{2\cdot N_1/2}\Big]\\ &\lesssim E\big[|\epsilon_1|^{N_1}\big]\cdot\sup_{n>0}\Big(\frac{1}{n}\sum_{i=1}^{n}|X_i|^{N_1}\Big) < \infty. \end{aligned} \tag{20}$$

The last boundedness of (20) follows from (8) and (10). 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 $a, b \ge 0$ and $N_2 = L(1-2\beta-\nu_2)^{-1} \ge 2$, we have

$$\begin{aligned}&\sup_{n>0} E\Big[\sup_{\theta\in\Theta}\Big(n^{1/2-\beta}\Big|\sum_{i=1}^{n}\frac{2}{n}\epsilon_i X_i[\theta-\theta_0] - C_n[\theta-\theta_0,\theta-\theta_0]+C_0[\theta-\theta_0,\theta-\theta_0]\Big|\Big)^{N_2}\Big]\\ &\lesssim \sup_{n>0}\Big(n^{-\beta N_2}E\Big[\Big|\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\epsilon_i X_i\Big|^{N_2}\Big]\Big) + \Big\{\sup_{n>0}\big(n^{1/2-\beta-\delta}n^{\delta}|C_n-C_0|\big)\Big\}^{N_2} < \infty. \end{aligned} \tag{21}$$

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

$$\sup_{n>0}\sup_{\theta\in\Theta}\Big[n^{1/2-\beta}\Big|\frac{1}{n}\sum_{j=1}^{p}\big\{\mathfrak{p}_n(\theta_j)-\mathfrak{p}_n(\theta_{0j})\big\}\Big|\Big]^{N_2} < \infty \tag{22}$$

since we assume (11). From (20)–(22), we conclude that [A5] holds. Therefore, the proof of (5) is complete because we established the PLDI (13). The latter claim of the theorem is trivial. □

Remark 1 We could deal with random design (X_i). Assume for simplicity that (X_i) and (ϵ_j) are independent. Then, in order to conclude (5), we need to change (9) and (10) into (23) and (24), respectively:

$$\exists\delta>0,\ \exists C_0>0:\text{constant},\ \forall k>0,\ \sup_{n>0}E\big[\big|n^{\delta}(C_n-C_0)\big|^{k}\big] < \infty. \tag{23}$$

$$\forall k > 0,\ \sup_{n>0} \sup_{i \le n} E\left[|X_{in}|^k\right] < \infty. \tag{24}$$

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

3 Examples

We will give some examples of the regularization term in (3) satisfying the conditions (11) and (12) in Theorem 2: sparse-bridge (Radchenko 2005), the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) and the Seamless-L_0 regularization (Dicker et al 2012). 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

$$Z_n(\theta) = Z_n(z, \rho) := \sum_{i=1}^{n} (Y_i - z^\top X_i^{(z)} - \rho^\top X_i^{(\rho)})^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|^\gamma, \tag{25}$$

where $\lambda_n \ge 0$ 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 (2005). He 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 $n^{-1} \max_{i \le n} |X_i|^2 \to 0$. Note that these conditions are satisfied with (8)–(10). Then, he proved the following results:

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

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

– 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 $\lambda_n / \sqrt{n} \to \lambda_0 \ge 0$;

(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 1 is derived from Theorem 2.

Corollary 1 *Assume that the linear regression model is* (2) *and the contrast function is* (25)*, where* $\lambda_n/n^{\gamma/2} \to \infty$ *and* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ *for* $\gamma \in (0, 1)$*. Suppose that we have* (8)–(10)*. Then, the PLDI* (5) *holds. In particular, the moment convergence* (6) *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 (11) and (12) in Theorem 2 for $\mathfrak{p}_n(\cdot) = \lambda_n|\cdot|^\gamma$. (11) 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 (12). 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 (12) holds for $\kappa = 1$, hence we obtain the PLDI (5). 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 $\mathcal{L}(\hat{v}_0) = N_{p_1}(-\lambda_0 B_0^{-1}\Upsilon, \sigma^2 B_0^{-1})$. □

Remark 2 Here, we briefly mention the case of the bridge-LSE $\hat{\theta}_n$ defined as the minimal point of the contrast function

$$Z_n(\theta) := \sum_{i=1}^n (Y_i - \theta^\top X_i)^2 + \lambda_n \sum_{j=1}^p |\theta_j|^\gamma, \tag{26}$$

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 (25), in (26) 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 (8)–(10). Then, Knight and Fu (2000) 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:

$$\hat{w}_0 = \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \operatorname{argmin}(V_0) \text{ if } \lambda_n / n^{(1\wedge\gamma)/2} \to \lambda_0 \geq 0,$$

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. 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 2014, Section 2) for details. □

3.2 SCAD

The SCAD-LSE (Fan and Li 2001) is defined as the minimum point of the contrast function (3), where

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

$\tau > 2$ is an additional tuning parameter. Let the minimizer be $\hat{\theta}_n = (\hat{z}_n, \hat{\rho}_n)$, and (8)–(10) hold. Then, under some conditions including

$$\lambda_n \to 0, \ \sqrt{n}\lambda_n \to \infty, \tag{27}$$

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

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

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. This meets (27). Now, we will show (11) and (12). First, we establish (11). 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 (11) holds. To ensure (12), we use

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

where $(x)_+ = \max(0, x)$. 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 (6), where $\hat{w}_0 = (0, \hat{v}_0)$ and $\mathcal{L}(\hat{v}_0) = N_{p_1}(0, \mathcal{I}_{p_1}^{-1}(\rho_0))$.

3.3 Seamless-L_0

The Seamless-L_0 regularization (Dicker et al 2012), 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 (8)–(10) hold. Then, under some conditions including

$$\lambda_n = O(1), \ \lambda_n\sqrt{n} \to \infty, \ \tau_n = O(n^{-3/2}), \tag{28}$$

Dicker et al (2012) 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 Sect. 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. This meets (28). Now, we will show (11) and (12) for $\mathfrak{p}_n(\cdot) =$

$(2n\lambda_n/\log 2)\log\{|\cdot|/(|\cdot|+\tau_n)+1\}$. (11) follows easily since $\mathfrak{p}_n/n^{1/2+\beta} \lesssim n^{1+\beta-1/2-1/2-\beta} = 1$. To ensure (12), we make use of the equation

$$|\log(1+x)-\log(1+x')| = \left|\int_0^1 \frac{\mathrm{d}s}{1+x'+(x-x')s}(x-x')\right|$$

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

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

Similarly, we get the same estimate for $a < 0$. As the results, it is possible to take the tuning parameters ensuring (6), 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 the referees for their helpful comments and suggestions. He also thanks Professor H. Masuda for his valuable comments.

References

Afendras, G., Markatou, M. (2015a). *Optimality of training/test size and resampling effectiveness of cross-validation estimators of the generalization error*. arXiv:1511.02980v1

Afendras, G., Markatou, M. (2015b). *U*niform integrability of the OLS estimators, and the convergence of their moments. arXiv:1511.02962v1

Chan, N. H., Ing, C.-K. (2011). Uniform moment bounds of Fisher's information with applications to time series. *The Annals of Statistics*, *39*(3), 1526–1550.

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

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.

Findley, D. F., Wei, C.-Z. (2002). AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *Journal of Multivariate Analysis*, *83*(2), 415–450.

Knight, K., Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378.

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

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

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.

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

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

Umezu, Y., Shimizu, Y., Masuda, H., Ninomiya, Y. (2015). *AIC for non-concave penalized likelihood method*. arXiv:1509.01688. (Submitted for publication).

van der Vaart, A. W. (1998). Asymptotic statistics. In *Cambridge Series in Statistical and Probabilistic Mathematics* (Vol. 3). Cambridge: Cambridge University Press.

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.