Stat Inference Stoch Process (2017) 20:387–397
DOI 10.1007/s11203-016-9146-0

Moment convergence of Z-estimators

Ilia Negri[1] · **Yoichi Nishiyama**[2]

Received: 31 May 2016 / Accepted: 11 July 2016 / Published online: 16 July 2016

Abstract The problem to establish the asymptotic distribution of statistical estimators as well as the moment convergence of such estimators has been recognized as an important issue in advanced theories of statistics. This problem has been deeply studied for M-estimators for a wide range of models by many authors. The purpose of this paper is to present an alternative and apparently simple theory to derive the moment convergence of Z-estimators. In the proposed approach the cases of parameters with different rate of convergence can be treated easily and smoothly and any large deviation type inequalities necessary for the same result for M-estimators do not appear in this approach. Applications to the model of i.i.d. observation, Cox's regression model as well as some diffusion process are discussed.

Keywords Asymptotic distribution · Method of moment estimators · Cox regression

1 Introduction

This paper is devoted to the convergence of moments for "Z-estimators", in other words, estimators that are the solutions to estimating equations.

For an illustration, let us consider the simplest case of i.i.d. data. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space, and let us be given a parametric family of probability densities $f(\cdot;\theta)$ with respect to μ, where $\theta \in \Theta \subset \mathbb{R}^d$. Let $X_1, X_2, \ldots$ be an independent sequence of $\mathcal{X}$-valued random variables from this parametric model. There are at least two ways to define the "maximum likelihood estimator (MLE)" in statistics. One way is to define it as the maximum point of the random function

✉ Ilia Negri
ilia.negri@unibg.it

[1] Department of Management, Information and Production Engineering, University of Bergamo, Viale Marconi, 5, 24044 Dalmine, BG, Italy

[2] Faculty of International Research and Education, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan

$$\theta \mapsto \mathbb{M}_n(\theta) = \frac{1}{n}\sum_{k=1}^{n} \log f(X_k;\theta),$$

while the other is to do it as the solution to the estimating equation

$$\mathbb{Z}_n(\theta) = 0, \quad \text{or, in another notation,} \quad \dot{\mathbb{M}}_n(\theta) = 0,$$

where $\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta)$ is the gradient vector of $\mathbb{M}_n(\theta)$. The former is a special case of "M-estimators", and the latter is that of "Z-estimators"; see van der Vaart and Wellner (1996) for these terminologies.

It is well known that the MLE $\widehat{\theta}_n$ is asymptotically normal: it holds for any *bounded* continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that

$$\lim_{n\to\infty} E[f(\sqrt{n}(\widehat{\theta}_n - \theta_0))] = E[f(I(\theta_0)^{-1/2}Z)],$$

where $I(\theta_0)$ is the Fisher information matrix and Z is a standard Gaussian random vector. Furthermore, it is important for some advanced theories in statistics, including asymptotic expansions and model selections, to extend this kind of results for *bounded* continuous functions f to that for any continuous function f with *polynomial growth*, that is, any continuous function f for which there exist some constants $C = C_f > 0$ and $q = q_f > 0$ such that

$$|f(x)| \le C(1 + ||x||)^q, \quad \forall x \in \mathbb{R}^d. \tag{1}$$

See the discussion in Yoshida (2011) for the importance of this problem.

Notice here that, when we have an asymptotic distribution result of an estimator, namely $R_n(\widehat{\theta}_n - \theta_0) \to^d L(\theta_0)$ where R_n is a (possibly, random) diagonal matrix and the limit random vector $L(\theta_0)$ is not necessarily Gaussian, it is sufficient for the generalisation to the case where f is a continuous function satisfying (1) to check that $||R_n(\widehat{\theta}_n - \theta_0)||$ is *asymptotically L_p-bounded* for some $p > q$, that is,

$$\limsup_{n\to\infty} E[||R_n(\widehat{\theta}_n - \theta_0)||^p] < \infty.$$

The study to provide some methods to obtain the moment convergence with polynomial order goes back to Ibragimov and Has'minskii (1981) who considered the MLEs and the Bayes estimators (as some special cases of M-estimators) in the general framework of the locally asymptotically normal models. It should be emphasised that one of the important merits of Ibragimov and Has'minskii's program is that the theory, based on the likelihood, automatically yields also the asymptotic efficiency of the estimators. In their main theorems, it was assumed that an *exponential type large deviation inequality* holds for the rescaled log-likelihood ratio random field. However, checking the assumption in terms of the large deviation inequality it has been not always easy. Although there are some successful works for some stochastic processes [see Kutoyants (1984) for general stochastic processes, Kutoyants (1994) for Poisson processes, and Kutoyants (2004) for ergodic diffusion processes], developing a general theory to establish the large deviation inequality was an open problem for many years. Finally in Yoshida (2011) it was pointed out that a *polynomial type large deviation inequality* is sufficient for the core part, i.e. the large deviation inequality, of the program presented in Ibragimov and Has'minskii (1981), and the main contribution is to have proved the (polynomial type) large deviation inequality with a good generality. Uchida and Yoshida (2012) applied the result in Yoshida (2011) to establish the moment convergence of some M-estimators in ergodic diffusion process models with an adjustment presented in Kessler's (1997). We also mention that in Nishiyama (2010) it is pointed out that the moment

convergence problem for M-estimators can be solved by using a maximal inequality instead of the large deviation inequalities, and that in Kato (2011) this type of approach is taken to deal with some bootstrap M-estimators.

In this paper, we consider the problem to prove the moment convergence of Z-estimators. In some context Z-estimators may be more natural than M-estimator, and in some cases Z-estimator cannot be derived from M-estimators, for example moment estimators are of such kind. By the way, as we have to assume that the random field (something like the log-likelihood) is differentiable, our framework is more restrictive than that for M-estimators. See Sect. 3.1. By contrast, the proofs becomes simpler because any large deviation type inequalities do not appear in our proof that is based on a combination of arguments involving the Hölder's and Minkowskii's inequalities.

Moreover it is possible to apply the result on Z-estimators to treat easily also the cases where the rates of convergence are different over the different components of θ. This is due to the fact that for Z-estimators we can multiply the gradient vector $\dot{\gamma}_n(\theta)$ of a contrast function $\gamma_n(\theta)$, where $\gamma_n(\theta)$ is typically the log-likelihood function, by a *matrix* R_n^{-2} to get a kind of law of large numbers, namely,

$$\dot{\mathbb{M}}_n(\theta) = R_n^{-2}\dot{\gamma}_n(\theta).$$

Typically, $R_n = \sqrt{n}I_d$ where I_d is the identity matrix, although a merit of our approach is that the diagonal components of R_n may be different in our framework. In contrast, in the framework of M-estimation theory the (scalar valued) contrast function $\gamma_n(\theta)$ with no assumption of differentiability has to be multiplied by a *scalar*. For ergodic diffusion process typically the rate of the parameters in drift coefficient is different from the rate of convergence of the parameters in the diffusion coefficient. As mentioned before, in Uchida and Yoshida (2012) the result in Yoshida (2011) is applied to establish the moment convergence of some M-estimators in ergodic diffusion process and they introduce some nuisance parameters in order to handle the components of different rates step by step. In Sect. 3.3 we present how the result can be obtained for the Z-estimators for the same model of ergodic diffusion process.

The rest of the paper is organised as follows. In the next Sect. 2 we define Z-estimator and we present our main result on convergence of moments for such estimators. In Sect. 3 we give some examples where the results presented in the previous section can be applied to have convergence of moments of appropriate Z-estimators. In particular in Sect. 3.1 the method of moments of i.i.d observation is presented. In Sect. 3.2 application to Cox's regression model is studied. In Sect. 3.3 examples to ergodic diffusion processes is given and finally in Sect. 3.4 an example on the estimation of volatility for diffusion processes is presented.

Before to close this section let us introduce some useful notations in what follows. The parameter space Θ is a bounded, open, convex subset of $\mathbb{R}^d$, where d is a fixed, positive integer. The word "vector" always means "d-dimensional real column vector", and the word "matrix" does "$d \times d$ real matrix". The Euclidean norm is denoted by $||v|| := \sqrt{\sum_{i=1}^d |v^{(i)}|^2}$ for a vector v where $v^{(i)}$ denotes the i-th component of v, and by $||A|| := \sqrt{\sum_{i,j=1}^d |A^{(i,j)}|^2}$ for a matrix A where $A^{(i,j)}$ denotes the (i, j)-component of A. Note that $||Av|| \leq ||A|| \cdot ||v||$ and $||AB|| \leq ||A|| \cdot ||B||$ for vector v and matrices A, B. The notations $v^\top$ and $A^\top$ denote the transpose. We use also the notation $A \circ B$ defined by $(A \circ B)^{(i,j)} := A^{(i,j)}B^{(i,j)}$ for two matrices A, B (the Hadamard product). We denote by I_d the identity matrix. The notations $\to^p$ and $\to^d$ mean the convergence in probability and the convergence in distribution, as $n \to \infty$, respectively.

2 Moment convergence of Z-estimators

Let Θ be a bounded, open, convex subset of $\mathbb{R}^d$. Let an $\mathbb{R}^d$-valued random function $\mathbb{Z}_n(\theta)$ of $\theta \in \Theta$ which is continuously differentiable with the gradient vector $\dot{\mathbb{Z}}_n(\theta)$, defined on a probability space $(\Omega, \mathcal{F}, P)$ that is common for all $n \in \mathbb{N}$. (However, it will be clear from our proofs that if the limit matrices $V(\theta_0)$ and $\dot{Z}(\theta)$ appearing below are non-random then the underlying probability spaces need not be common for all $n \in \mathbb{N}$.)

An important special case is $\mathbb{Z}_n(\theta)$ given as the gradient vector $\dot{\mathbb{M}}_n(\theta)$ of a rescaled contrast function $\mathbb{M}_n(\theta) = R_n^{-2}\gamma_n(\theta)$ of $\theta \in \Theta$ which is twice continuously differentiable with the gradient vector $\dot{\mathbb{M}}_n(\theta)$ and the Hessian matrix $\ddot{\mathbb{M}}_n(\theta)$, where R_n be a (possibly, random) diagonal matrix whose diagonal components are positive; that is, defining Q_n by $Q_n^{(i,j)} = (R_n^{(i,i)} R_n^{(j,j)})^{-1}$, put

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = R_n^{-2}\dot{\gamma}_n(\theta) \quad \text{and} \quad \dot{\mathbb{Z}}_n(\theta) = \ddot{\mathbb{M}}_n(\theta) = Q_n \circ \ddot{\gamma}_n(\theta). \tag{2}$$

Remark In the typical cases, $R_n = \sqrt{n} I_d$ and $Q_n = n^{-1}\mathbf{1}$, where $\mathbf{1}$ denotes the matrix whose all components are 1. In same specific models, such as branching process for example, the rate matrix R_n for the estimators of the offspring distribution may depend on θ_0.

Turning back to the general setup, we shall state a theorem to give an asymptotic representation for Z-estimators.

Let us introduce the following conditions.

[Z1] *Suppose there exists a sequence of matrices $V_n(\theta_0)$ which are regular almost surely such that for any sequence of Θ-valued random vectors $\widetilde{\theta}_n$ converging in probability to θ_0,*

$$\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) - (-V_n(\theta_0)) \to^p 0.$$

[Z2] *Suppose that*

$$(R_n\mathbb{Z}_n(\theta_0), V_n(\theta_0)) \to^d (L(\theta_0), V(\theta_0)),$$

where $R_n$be a (possibly, random) diagonal matrix whose diagonal components are positive, $L(\theta_0)$ is a random vector, and $V(\theta_0)$ is a random matrix which is regular almost surely.

Remark In condition [Z2] we do not assume that $V(\theta_0)$ and $L(\theta_0)$ are independent.

Although the following result is not really novel, we will give a full (and short) proof for references.

Theorem 2.1 *Let an $\mathbb{R}^d$-valued random function $\mathbb{Z}_n(\theta)$ of $\theta \in \Theta$ which is continuously differentiable with the gradient vector $\dot{\mathbb{Z}}_n(\theta)$ be given. Suppose that condition [Z1] and [Z2] hold true.*

Then, for any sequence of Θ-valued random vectors $\widehat{\theta}_n$ which converges in probability to θ_0 and satisfies that $||R_n\mathbb{Z}_n(\widehat{\theta}_n)|| = o_P(1)$, it holds that

$$\begin{aligned} R_n(\widehat{\theta}_n - \theta_0) &= V_n(\theta_0)^{-1} R_n \mathbb{Z}_n(\theta_0) + o_P(1) \\ &\to^d V(\theta_0)^{-1} L(\theta_0). \end{aligned}$$

Remark Usually the matrices $V_n(\theta) = -\dot{\mathbb{Z}}_n(\theta)$.

In Theorem 2.1 the consistency of the sequence of Z-estimators $\widehat{\theta}_n$ has been assumed. A method to show this property will be given in Lemma 2.2 below, whose proof is omitted because it can be proved exactly in the same way as Theorems 5.7 and 5.9 of van der Vaart (1998).

Lemma 2.2 *Suppose that for some $\theta_0 \in \Theta$, it holds that*

$$\sup_{\theta\in\Theta} ||\mathbb{Z}_n(\theta) - Z_{\theta_0}(\theta)|| \to^p 0,$$

where the random field $\theta \rightsquigarrow Z_{\theta_0}(\theta)$ of the limit satisfies that

$$\inf_{\theta:||\theta-\theta_0||>\varepsilon} ||Z_{\theta_0}(\theta)|| > 0 = ||Z_{\theta_0}(\theta_0)||, \quad \textit{almost surely,} \quad \forall \varepsilon > 0.$$

Then, for any sequence of Θ-valued random vectors $\widehat{\theta}_n$ such that $||\mathbb{Z}_n(\widehat{\theta}_n)|| = o_P(1)$, it holds that $\widehat{\theta}_n \to^p \theta_0$.

In the following we define R_n^r for any $r \in \mathbb{R}$ as $R_n^r = diag[(R_n^{(1,1)})^r, \ldots, (R_n^{(d,d)})^r]$. Let some constants $p \geq 1$ and $a, b > 1$ such that $\frac{1}{a} + \frac{1}{b} = 1$ be given. Let us introduce the following conditions.

[Z3] *Suppose that for some $\theta_0 \in \Theta$,*

$$||R_n\mathbb{Z}_n(\theta_0)|| \text{ is asymptotically } L_{pa}\text{-bounded.} \tag{3}$$

[Z4] *Suppose that there exist a constant $\gamma \in (0, 1]$ and some random matrices $\dot{Z}_{\theta_0}(\theta)$ indexed by $\theta \in \Theta$ such that*

$$\lim_{n\to\infty} E\left[\sup_{\theta\in\Theta} ||R_n^{\gamma}(\dot{\mathbb{Z}}_n(\theta) - \dot{Z}_{\theta_0}(\theta))||^{pa/\gamma}\right] = 0. \tag{4}$$

[M1] *There exists a random matrix $J(\theta_0)$ which is positive definite almost surely such that $\dot{Z}_{\theta_0}(\theta) \leq -J(\theta_0)$ for all $\theta \in \Theta$, almost surely, and that $E[||J(\theta_0)^{-1}||^{pb/\gamma}] < \infty$.*

[M2] *$E[\sup_{\theta\in\Theta} ||\dot{Z}_{\theta_0}(\theta)^{-1}||^{pb/\gamma}] < \infty$, where the random matrices $\dot{Z}_{\theta_0}(\theta)$'s are assumed to be regular almost surely.*

Now, we give a theorem to establish the moment convergence of Z-estimators, which is the main result in this paper.

Theorem 2.3 *Let an $\mathbb{R}^d$-valued random function $\mathbb{Z}_n(\theta)$ of $\theta \in \Theta$ which is continuously differentiable with the gradient vector $\dot{\mathbb{Z}}_n(\theta)$ be given. Suppose that conditions [Z3] and [Z4] hold true. Suppose further that either of the conditions [M1] or [M2] is satisfied.*

Then, for any sequence of Θ-valued random vectors $\widehat{\theta}_n$ such that $||R_n\mathbb{Z}_n(\widehat{\theta}_n)||$ is asymptotically L_{pa}-bounded, it holds that $||R_n(\widehat{\theta}_n - \theta_0)||$ is asymptotically L_p-bounded. Therefore, in this situation, whenever we also have that $R_n(\widehat{\theta}_n - \theta_0) \to^d G(\theta_0)$ where $G(\theta_0)$ is a random vector, it holds for any continuous function f satisfying (1) for $q \in (0, p)$ that

$$\lim_{n\to\infty} E[f(R_n(\widehat{\theta}_n - \theta_0))] = E[f(G(\theta_0))],$$

where the limit is also finite.

In the last theorem we can observe that when the second condition in [M1] is satisfied with $||J(\theta_0)||^{-1}$ which is bounded or the first condition in [M2] is satisfied with $\sup_{\theta\in\Theta} ||\dot{Z}_{\theta_0}(\theta)^{-1}||$ which is bounded, the constant a appearing in the above claim may be replaced by 1.

Remark The crucial point in the course of applying this theorem is to check the condition (4) together with [M1] or [M2]. This is clearly satisfied for moment estimators described in Example 3.1.

Remark If the matrix $\dot{\mathbb{Z}}_n(\theta)$ is symmetric than condition [M1] can be satisfied if the smallest eigenvalues is L_p-bounded. If the matrix $\dot{\mathbb{Z}}_n(\theta)$ is symmetric and random than condition [M1] can be satisfied if all the eigenvalues are L_p-bounded.

Remark The condition [M1] above is corresponding to the case $\rho = 2$ of the conditions [A3] and [A5] in Yoshida (2011), which are, rewritten with our notation

$$M_{\theta_0}(\theta) - M_{\theta_0}(\theta_0) \leq -\chi(\theta_0)||\theta - \theta_0||^{\rho}, \quad \forall \theta \in \Theta,$$

where $M_{\theta_0}(\theta)$ denotes the "limit" of $\mathbb{M}_n(\theta)$, and high order moment conditions on the positive random variable $\chi(\theta_0)^{-1}$.

Remark Condition (3) for many models can be checked by applying the Burkholder inequality [see Theorem 26.12 in Kallenberg (2002)]. See Example 3.1.

Proof of Theorem 2.1. Recalling (2), it follows from the Taylor expansion that

$$(R_n\mathbb{Z}_n(\widehat{\theta}_n))^{(i)} = (R_n\mathbb{Z}_n(\theta_0))^{(i)} + (\dot{\mathbb{Z}}_n(\widetilde{\theta}_n)R_n(\widehat{\theta}_n - \theta_0))^{(i)}, \quad i = 1, \ldots, d. \tag{5}$$

So we have

$$R_n(\widehat{\theta}_n - \theta_0) = A_n + B_n R_n(\widehat{\theta}_n - \theta_0), \tag{6}$$

where

$$\begin{aligned} A_n &= V_n(\theta_0)^{-1} R_n(\mathbb{Z}_n(\theta_0) - \mathbb{Z}_n(\widehat{\theta}_n)), \\ B_n &= V_n(\theta_0)^{-1}(\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) + V_n(\theta_0)), \end{aligned}$$

and $\widetilde{\theta}_n$ is a random vector on the segment connecting θ_0 and $\widehat{\theta}_n$. It follows from the extended continuous mapping theorem that $V_n(\theta_0)^{-1} \to^p V(\theta_0)^{-1}$ [e.g., Theorem 1.11.1 of van der Vaart and Wellner (1996)], thus we have $||A_n|| = O_P(1)$ and $||B_n|| = o_P(1)$. It therefore holds that

$$||R_n(\widehat{\theta}_n - \theta_0)|| \leq O_P(1) + o_P(1) \cdot ||R_n(\widehat{\theta}_n - \theta_0)||,$$

which implies that $||R_n(\widehat{\theta}_n - \theta_0)|| = O_P(1)$. Hence, going back to (6) we obtain

$$R_n(\widehat{\theta}_n - \theta_0) = A_n + o_P(1) = V_n(\theta_0)^{-1} R_n \mathbb{Z}_n(\theta_0) + o_P(1).$$

The last claim is also a consequence of the extended continuous mapping theorem. The proof is finished. □

Proof of Theorem 2.3. We will give a proof for the case where [M1] is assumed. The proof for the case where [M2] is assumed is similar (and simpler), so it is omitted.

Due to (5) again, we have

$$R_n(\widehat{\theta}_n - \theta_0) = C_n + \left(D_n^{(1)} + D_n^{(2)}\right) R_n(\widehat{\theta}_n - \theta_0),$$

where

$$\begin{aligned} C_n &= J(\theta_0)^{-1} R_n(\mathbb{Z}_n(\theta_0) - \mathbb{Z}_n(\widehat{\theta}_n)), \\ D_n^{(1)} &= J(\theta_0)^{-1}(\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) - \dot{Z}_{\theta_0}(\widetilde{\theta}_n)), \\ D_n^{(2)} &= J(\theta_0)^{-1}(\dot{Z}_{\theta_0}(\widetilde{\theta}_n) + J(\theta_0)), \end{aligned}$$

where $\widetilde{\theta}_n$ is a random vector on the segment connecting θ_0 and $\widehat{\theta}_n$.

From now on, we consider the case $\gamma \in (0, 1)$; the proof for the case $\gamma = 1$ is easier, and it is omitted. Since $-D_n^{(2)}$ is non-negative definite almost surely, it follows from Minkowskii's and Hölder's inequalities that

$$\begin{aligned}
&(E[||R_n(\widehat{\theta}_n - \theta_0)||^p])^{1/p} \\
&\quad \le (E[||(I_d - D_n^{(2)})R_n(\widehat{\theta}_n - \theta_0)||^p])^{1/p} \\
&\quad \le (E[||C_n||^p])^{1/p} + \left(E\left[||R_n^{\gamma} D_n^{(1)}||^{p/\gamma}\right]\right)^{\gamma/p} \left(E\left[||R_n^{1-\gamma}(\widehat{\theta}_n - \theta_0)||^{p/(1-\gamma)}\right]\right)^{(1-\gamma)/p} \\
&\quad \le O(1) + o(1) \cdot \left(E\left[||R_n^{1-\gamma}(\widehat{\theta}_n - \theta_0)||^{p/(1-\gamma)}\right]\right)^{(1-\gamma)/p},
\end{aligned}$$

where we have used Hölder's inequality again to get

$$E[||C_n||^p] \le \left(E\left[||J(\theta_0)^{-1}||^{pb}\right]\right)^{1/b} (E[||R_n(\mathbb{Z}_n(\theta_0) - \mathbb{Z}_n(\widehat{\theta}_n))||]^{pa})^{1/a}$$

and

$$E[||R_n^{\gamma} D_n^{(1)}||^{p/\gamma}] \le \left(E\left[||J(\theta_0)^{-1}||^{pb/\gamma}\right]\right)^{1/b} (E[||R^{\gamma}(\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) - \dot{Z}_{\theta_0}(\widetilde{\theta}_n))||]^{pa/\gamma})^{1/a};$$

if $||J(\theta_0)||^{-1}$ is bounded, we can get this kind of bound with $a = 1$.

Notice that

$$\begin{aligned}
||R_n^{1-\gamma}(\widehat{\theta}_n - \theta_0)||^{1/(1-\gamma)} &\le \sqrt{d^{(1/(1-\gamma))-1} \sum_{i=1}^{d} |R_n^{(i,i)}|^2 |\widehat{\theta}_n^{(i)} - \theta_0^{(i)}|^{2/(1-\gamma)}} \\
&\le ||R_n(\widehat{\theta}_n - \theta_0)|| \cdot d^{1/(2-2\gamma)} \cdot |\mathcal{D}(\Theta)|^{\gamma/(1-\gamma)},
\end{aligned}$$

where $\mathcal{D}(\Theta)$ denotes the diameter of Θ. So we obtain

$$\begin{aligned}
(E[||R_n(\widehat{\theta}_n - \theta_0)||^p])^{1/p} &\le O(1) + o(1) \cdot (E[||R_n(\widehat{\theta}_n - \theta_0)||^p])^{(1-\gamma)/p} \\
&\le O(1) + o(1) \cdot (E[||R_n(\widehat{\theta}_n - \theta_0)||^p] \vee 1)^{1/p},
\end{aligned}$$

which yields that

$$E[||R_n(\widehat{\theta}_n - \theta_0)||^p] \le O(1) + o(1) \cdot E[||R_n(\widehat{\theta}_n - \theta_0)||^p].$$

Therefore, $||R_n(\widehat{\theta}_n - \theta_0)||$ is asymptotically L_p-bounded. □

3 Examples

In this section we give some examples where the results presented in the previous section can be applied to have convergence of moments of appropriate Z-estimators.

3.1 Moment estimators

Let $X, X_1, X_2, \ldots$ be an i.i.d. sample from a distribution P_θ on $(\mathcal{X}, \mathcal{A})$, where $\theta \in \Theta \subset \mathbb{R}^d$. Let $\psi^{(1)}, \ldots, \psi^{(d)}$ be measurable functions on $\mathcal{X}$. Define

$$\mathbb{Z}_n(\theta) = \frac{1}{n}\sum_{k=1}^{n}(\psi^{(1)}(X_k) - e^{(1)}(\theta), \ldots, \psi^{(d)}(X_k) - e^{(d)}(\theta))^{\top} = \frac{1}{n}\sum_{k=1}^{n}(\psi(X_k) - e(\theta)),$$

where $e(\theta) = (E_\theta[\psi^{(1)}(X)], \dots E_\theta[\psi^{(d)}(X)])^\top$. The solution $\hat{\theta}_n$ of the system of equations $\mathbb{Z}_n(\hat{\theta}_n) = 0$ is called method of moments estimator. If e is one-to-one then the moment estimator is uniquely determined as $\hat{\theta}_n = e^{-1}(\frac{1}{n}\sum_{k=1}^n(\psi(X_k)))$.

The results of the previous Sect. 2 can be applied to these estimating functions whose derivative matrix is $\dot{\mathbb{Z}}_n(\theta) = \{\dot{\mathbb{Z}}_n^{(i,j)}(\theta)\}_{(i,j)\in\{1,\dots,d\}^2}$, where $\dot{\mathbb{Z}}_n^{(i,j)}(\theta) = \frac{\partial}{\partial\theta_j}\mathbb{Z}_n^{(i)}(\theta) = -\frac{\partial}{\partial\theta_j}e^{(i)}(\theta)$. Let us define the matrix $V(\theta) = -\dot{Z}(\theta) = \frac{\partial}{\partial\theta_j}e^{(i)}(\theta)$. Let us suppose that $V(\theta)$ is invertible and that $E_{\theta_0}[||\psi(X)||^{2p}] < \infty$. Then condition (3) is satisfied with $a = 2$, $R_n = \sqrt{n}I_d$ by applying the Burkholder inequality [see Theorem 26.12 in Kallenberg (2002)] to the martingale $M_n = \frac{1}{\sqrt{n}}\sum_{k=1}^n(\psi(X_k) - e(\theta))$. Conditions (4) and [M1] are trivially satisfied with $\gamma = 1$ and $b = 2$ because $\dot{\mathbb{Z}}_n(\theta) = -\dot{Z}(\theta) = V(\theta)$ is non random.

Moreover denote with $L(\theta_0)$ a Gaussian d-vector with 0 mean and covariance matrix given by Σ_{θ_0} with $\Sigma_{\theta_0}^{(i,j)} = \left\{E_{\theta_0}[(\psi^{(i)}(X_k) - e^{(i)}(\theta_0))(\psi^{(j)}(X_k) - e^{(j)}(\theta_0))]\right\}$. Then Theorem 2.3 holds with $G(\theta_0) = V^{-1}(\theta_0)L(\theta_0)$.

3.2 Cox's regression model

Let a sequence of counting processes $t \rightsquigarrow N_t^k, k = 1, 2, \dots$, which do not have simultaneous jumps, be observed the time interval $[0, T]$. Suppose that $t \rightsquigarrow N_t^k$ has the intensity

$$\lambda_t^k(\theta) = \alpha(t)e^{\theta^\top Z_t^k}Y_t^k,$$

where the baseline hazard function α which is common for all k's is non-negative and satisfies that $\int_0^T \alpha(t)dt < \infty$, the random process $t \rightsquigarrow Z_t^k$ is a vector valued bounded covariate for the individual k, and the random process $t \rightsquigarrow Y_t^k$ is given by

$$Y_t^k = \begin{cases} 1, & \text{if the individual } k \text{ is observed at time t,} \\ 0, & \text{otherwise.} \end{cases}$$

Let $\theta \in \Theta \subset \mathbb{R}^d$, where Θ is a compact set. This model was introduced in Cox (1972), and its asymptotic theory was developed in Andersen and Gill (1982). Assuming bounded covariate is often done in applications. We introduce

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = \frac{1}{n}\dot{\gamma}_n(\theta),$$

where

$$\gamma_n(\theta) = \sum_{k=1}^n \int_0^T \left(\theta^\top Z_t^k - \log S_t^{n,0}(\theta)\right) dN_t^k$$

with

$$S_t^{n,0}(\theta) = \sum_{k=1}^n e^{\theta^\top Z_t^k}Y_t^k.$$

The rate matrix is $R_n = \sqrt{n}I_d$.

Let us define

$$S_t^{n,1}(\theta) = \sum_{k=1}^{n} Z_t^k e^{\theta^\top Z_t^k} Y_t^k,$$
$$S_t^{n,2}(\theta) = \sum_{k=1}^{n} \left(Z_t^k\right)^\top Z_t^k e^{\theta^\top Z_t^k} Y_t^k,$$

Let us introduce the following conditions:

There exists a constant $\gamma > 0$ such that

$$\sup_{\theta\in\Theta} \int_0^T \frac{\sqrt{n}^\gamma}{n} \left| \frac{S_t^{n,l}(\theta) S_t^{n,m}(\theta)^\top}{S_t^{n,0}(\theta)} - \frac{\mathcal{S}_t^l(\theta)\mathcal{S}_t^m(\theta)^\top}{\mathcal{S}_t^0(\theta)} \right| dt, \quad (l,m) = (2,0),(1,1) \tag{7}$$

is L_p-bounded for any $p \geq 1$. The limits $t \rightsquigarrow \mathcal{S}_t^l$ are some stochastic processes [see Andersen and Gill (1982) who assumed that $\mathcal{S}^l$'s are not random].

Moreover let us suppose that the inverse of all the eigenvalues of the matrix $V(\theta_0)$ where

$$V(\theta) = \int_0^T \frac{\mathcal{S}_t^0(\theta)\mathcal{S}_t^2(\theta) - \mathcal{S}_t^1(\theta)\mathcal{S}_t^1(\theta)^\top}{\mathcal{S}_t^0(\theta)^2} \mathcal{S}_t^0(\theta_0)\alpha(t)dt$$

are L_p-bounded.

Let us now define

$$\mathbb{Z}_{\theta_0}(\theta) = \int_0^T \left(\frac{\mathcal{S}_t^1(\theta_0)}{\mathcal{S}_t^0(\theta_0)} - \frac{\mathcal{S}_t^1(\theta)}{\mathcal{S}_t^0(\theta)} \right) \mathcal{S}_t^0(\theta_0)\alpha(t)dt,$$
$$V_n(\theta_0) = \frac{1}{n} \int_0^T \frac{S_t^{n,0}(\theta_0) S_t^{n,2}(\theta_0) - S_t^{n,1}(\theta_0) S_t^{n,1}(\theta_0)^\top}{S_t^{n,0}(\theta_0)} \alpha(t)dt,$$

Now observing that $\dot{\gamma}_n(\theta)$ can be rewritten in θ_0 as

$$\dot{\gamma}_n(\theta_0) = \sum_{k=1}^{n} \int_0^T \left(Z_t^k - \frac{S_t^{n,1}(\theta_0)}{S_t^{n,0}(\theta_0)} \right) (dN_t^k - \alpha(t) e^{\theta^\top Z_t^k} Y_t^k dt),$$

(we indeed add a quantity that is 0), then condition (3) with $a = 2$ can be proved applying the Burkholder inequality [see Theorem 26.12 in Kallenberg (2002)] to the martingale $M_n = R_n \mathbb{Z}_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \int_0^T \left(Z_t^k - \frac{S_t^{n,1}(\theta_0)}{S_t^{n,0}(\theta_0)} \right) (dN_t^k - \alpha(t) e^{\theta^\top Z_t^k} Y_t^k dt)$.

To prove condition (4) let us consider first $\dot{\mathbb{Z}}_n(\theta) + V_n(\theta)$. We have

$$\dot{\mathbb{Z}}_n(\theta) = -\frac{1}{n} \sum_{k=1}^{n} \int_0^T \left(\frac{S_t^{n,0}(\theta_0) S_t^{n,2}(\theta_0) - S_t^{n,1}(\theta_0) S_t^{n,1}(\theta_0)^\top}{S_t^{n,0}(\theta_0)^2} \right) dN_t^k$$

and $\dot{\mathbb{Z}}_n(\theta) + V_n(\theta)$ is a martingale for any θ. Now let us consider $V(\theta) - V_n(\theta)$. $\sup_\theta |V(\theta) - V_n(\theta)|$ can be bounded by the triangular inequality and condition (7).

Finally condition [M1] follow from the condition of boundedness of eigenvalues of the matrix $V(\theta_0)$.

3.3 Ergodic diffusion process

Let $I = (l, r)$, where $-\infty \le l < r \le \infty$, be given. Let us consider an I-valued diffusion process $t \rightsquigarrow X_t$ which is the unique strong solution to the stochastic differential equation (SDE)

$$X_t = X_0 + \int_0^t S(X_s; \alpha)ds + \int_0^t \sigma(X_s; \beta)dW_s,$$

where $s \rightsquigarrow W_s$ is a standard Wiener process. The first d_A-components $\alpha \in \Theta_A \subset \mathbb{R}^{d_A}$ of the parameter $\theta = (\alpha^\top, \beta^\top)^\top$ is involved in the drift coefficient, and the latter d_B-components $\beta \in \Theta_B \subset \mathbb{R}^{d_B}$ is in the diffusion coefficient. We are supposed to be able to observe the process X at discrete time grids $0 = t_0^n < t_1^n < \cdots < t_n^n$, and we shall consider the asymptotic scheme $n\Delta_n^2 \to 0$ and $t_n^n \to \infty$ as $n \to \infty$, where

$$\Delta_n = \max_{1 \le k \le n} |t_k^n - t_{k-1}^n|,$$

and

$$\sum_{k=1}^n \left| \frac{|t_k^n - t_{k-1}^n|}{t_n^n} - \frac{1}{n} \right| \to 0, \quad \text{as } n \to \infty. \tag{8}$$

The problem to establish the moment convergence for M-estimators for ergodic diffusion processes where X is a multi-dimensional diffusion process, was considered in Yoshida (2011). In Uchida and Yoshida (2012) the assumption $n\Delta_n^2 \to 0$ was relaxed up to $n\Delta_n^a \to 0$, where $a \ge 2$ is a constant depending on the smoothness of the model, by using a method presented in Kessler's (1997). Their arguments consist of plural steps in order to handle the parameters α and β, whose rates of convergence are different, separately. The advantages of taking Z-estimator makes possible to treat both parameters simultaneously and the convergence of moments can be obtained easily with the following setting. Define

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = R_n^{-2}\dot{\gamma}_n(\theta),$$

where

$$\gamma_n(\theta) = -\sum_{k: t_{k-1}^n \le t_n^n} \left\{ \log \sigma(X_{t_{k-1}^n}; \beta) + \frac{|X_{t_k^n} - X_{t_{k-1}^n} - S\left(X_{t_{k-1}^n}; \alpha\right)|t_k^n - t_{k-1}^n||^2}{2\sigma\left(X_{t_{k-1}^n}; \beta\right)^2 |t_k^n - t_{k-1}^n|} \right\}$$

and R_n is the diagonal matrix such that $R_n^{(i,i)}$ is $\sqrt{t_n^n}$ for $i = 1, \ldots, d_A$ and $\sqrt{n}$ for $i = d_A + 1, \ldots, d$ with $d = d_A + d_B$. With the above setting convergence of moments for the Z-estimators of α and β can be obtained easily and in a compact form. Here we omit the details because the results are the same as in Uchida and Yoshida (2012).

3.4 Volatility of diffusion process

Let $I = (l, r)$, where $-\infty \le l < r \le \infty$, be given. Let us consider an I-valued diffusion process $t \rightsquigarrow X_t$ which is the unique strong solution to the SDE

$$X_t = X_0 + \int_0^t S(X_s)ds + \int_0^t \sigma(X_s; \theta)dW_s,$$

where $s \rightsquigarrow W_s$ is a standard Wiener process. Here, the drift coefficient $S(\cdot)$ is treated as an unknown nuisance function. We are supposed to be able to observe the process X at discrete

time grids $0 = t_0^n < t_1^n < \cdots < t_n^n = T < \infty$, and we shall consider the asymptotic scheme (8).

The same model was considered in Uchida and Yoshida (2013) where they prove the convergence of moments for the quasi likelihood estimator. The convergence of moment for the Z-estimator of the parameter θ can be proved also with the following setting.

We introduce

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = \frac{1}{n}\dot{\gamma}_n(\theta),$$

where

$$\gamma_n(\theta) = -\sum_{k:t_{k-1}^n \le t_n^n} \left\{ \log \sigma\left(X_{t_{k-1}^n}; \theta\right) + \frac{|X_{t_k^n} - X_{t_{k-1}^n}|^2}{2\sigma\left(X_{t_{k-1}^n}; \theta\right)^2 |t_k^n - t_{k-1}^n|} \right\}.$$

The rate matrix is given by $R_n = \sqrt{n} I_d$. With the above setting convergence of moments for the Z-estimators of θ can be obtained easily and in a compact form. Here we omit the details because the results are the same as in Uchida and Yoshida (2013).

Acknowledgements The authors thank an anonymous referee for her or his helpful comments. This work was supported by Italian MIUR, Grant 2009 (I.N.) and by Grant-in-Aid for Scientific Research (C), 24540152, from Japan Society for the Promotion of Science (Y.N.).

References

Andersen PK, Gill RD (1982) Cox's regression models for counting processes: a large sample study. Ann Stat 10:1100–1120

Cox DR (1972) Regression models and life-tables (with discussion). J R Stat Soc B 34:187–220

Ibragimov IA, Has'minskii RZ (1981) Statistical estimation: asymptotic theory. Springer, New York

Kallenberg O (2002) Foundations of modern probability, 2nd edn. Springer, New York

Kato K (2011) A note on moment convergence of bootstrap M-estimators. Stat Decis 28:51–61

Kessler M (1997) Estimation of an ergodic diffusion from discrete observations. Scand J Stat 24:211–229

Kutoyants YA (1984) Parameter estimation for stochastic processes. Heldermann, Berlin

Kutoyants YA (1994) Identification of dynamical systems with small noise. Kluwer Academic Publishers, Dordrecht

Kutoyants YuA (2004) Statistical inference for ergodic diffusion processes. Springer, London

Nishiyama Y (2010) Moment convergence of M-estimators. Stat Neerl 64:505–507

Uchida M, Yoshida N (2012) Adaptive estimation of an ergodic diffusion process based on sampled data. Stoch Process Appl 122:2885–2924

Uchida M, Yoshida N (2013) Quasi likelihood analysis of volatility and nondegeneracy of statistical random field. Stoch Process Appl 123:2851–2876

van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge

van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes: with applications to statistics. Springer, New York

Yoshida N (2011) Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. Ann Inst Stat Math 63:431–479