Z-ESTIMATORS INDEXED BY OBJECTIVE FUNCTIONS

François Portier [1]

[1] *20 rue du roman pays, 1348 Louvain-La-Neuve, Belgium,*
email: francois.portier@gmail.com

Résumé. On étudie la convergence de Z-estimateurs $\hat{\theta}(\eta)$ pour lesquels la fonction objective dépend d'un paramètre η appartenant à un espace de Banach $\mathcal{H}$. On démontre la consistence uniforme sur $\mathcal{H}$, la convergence faible dans $l^\infty(\mathcal{H})$ et la validité du bootstrap. Lorsque η est un paramètre de "tuning" ayant pour valeur optimale η_0, on donne des conditions pour qu'un estimateur $\hat{\eta}$ puisse être remplacé par η_0 sans changer la variance asymptotique. Ces conditions ne demandent pas de vitesse particulière concernant la convergence de $\hat{\eta}$ vers η_0. De manière similaire on montre que le bootstrap de $\hat{\theta}(\hat{\eta})$ est valide même sans effectuer un bootstrap de $\hat{\eta}$. On s'intéresse à plusieurs applications et on étudie plus en details le cas où η est la fonction de poids d'une régression pondérée. Cette nouvelle approche permet d'obtenir des conditions générales quant à la procédure d'estimation des poids optimaux. La précision de différentes procédures est évaluée par simulation.

Mots-clés. Z-estimateurs, efficacité, bootstrap, processus empirique, régression pondérée.

Abstract. We study the convergence of Z-estimators $\hat{\theta}(\eta)$ for which the objective function depends on a parameter η that belongs to a Banach space $\mathcal{H}$. Our results include uniform consistency over $\mathcal{H}$, weak convergence in $l^\infty(\mathcal{H})$ and the validity of the bootstrap. Furthermore when η is a tuning parameter optimally selected at η_0, we give conditions under which an estimated $\hat{\eta}$ can be replaced by η_0 without affecting the asymptotic variance. Interestingly, these conditions are free from any rate of convergence of $\hat{\eta}$ to η_0. A related feature is that the estimator $\hat{\theta}(\hat{\eta})$ is bootstrapped without the need of bootstrapping $\hat{\eta}$. We highlight several applications of our results and we study in detail the case where η is the function of weight in weighted regression. Our alternative treatment allows to obtain new general conditions related to the estimation procedure of the optimal weights. Small sample performances of different procedures are discussed through simulations.

Keywords. Z-estimators, efficiency, bootstrap, empirical process, weighted regression.

1 Asymptotic equicontinuity to achieve efficiency

Let P be a probability measure defined on the measurable space $(\mathcal{Z}, \mathcal{A})$ and $(Z_1, \ldots, Z_n)$ be i.i.d. random elements with law P. For a measurable function $f : \mathcal{Z} \to \mathbb{R}$, we define

$$Pf = \int f dP, \qquad \mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(Z_i), \qquad \mathbb{G}_n f = n^{1/2}(\mathbb{P}_n - P)f,$$

where the latter one is called the empirical process. Considering the estimation of a Euclidean parameter $\theta_0 \in \Theta \subset \mathbb{R}^p$, we denote by $(\mathcal{H}, \|\cdot\|)$ a Banach space and we let $\{\hat{\theta}(\eta),\ \eta \in \mathcal{H}\}$ be a collection of estimators based on $(Z_1, \ldots, Z_n)$. Suppose furthermore that there exists $\eta_0 \in \mathcal{H}$ such that $\hat{\theta}(\eta_0)$ is efficient within this collection, i.e. $\hat{\theta}(\eta_0)$ has the smallest variance among the estimators of the class. Such a situation arise in many fields of the statistics: for instance, η can be the cut-off parameter in Huber robust regression, or η might as well equal a function of weights in heteroscedastic regression (see later for more details and examples). Unfortunately, η_0 is generally unknown from us since it certainly depends on the model P. Usually, one is restricted to first estimate η_0 by, say $\hat{\eta}$, and then to compute the estimator of θ_0: $\hat{\theta}(\hat{\eta})$. Whereas it is reasonable to believe that $\hat{\theta}(\hat{\eta})$ is not a too bad estimator of θ_0, it turns out that, in many different situations, $\hat{\theta}(\hat{\eta})$ actually achieves the efficiency bound of the collection (see for instance Newey and McFadden (1994) or the examples later). This is even more surprising as soon as we know that the accuracy of $\hat{\eta}$ estimating η_0 does not matter, provided its consistency. A paradigm that encompasses the latter facts can be developed *via* the notion of asymptotic equicontinuity. Define the process $\eta \mapsto \mathbb{Z}_n(\eta) = \sqrt{n}(\hat{\theta}(\eta) - \theta_0)$ and suppose that it lies in $l^\infty(\mathcal{H})$, it is asymptotically equicontinuous if for any $\epsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to +\infty} P\Big(\sup_{\|\eta_1 - \eta_2\| < \delta} |\mathbb{Z}_n(\eta_1) - \mathbb{Z}_n(\eta_2))| > \epsilon \Big) = 0. \tag{1}$$

Clearly, for $\hat{\theta}(\hat{\eta})$ to be efficient it suffices that $\sqrt{n}(\hat{\theta}(\hat{\eta}) - \hat{\theta}(\eta_0)) = \mathbb{Z}_n(\hat{\eta}) - \mathbb{Z}_n(\eta_0)$ goes to 0 in probability. This is true if, in addition to (1), the following hold

$$P(\hat{\eta} \in \mathcal{H}) \longrightarrow 1 \qquad \text{and} \qquad \|\hat{\eta} - \eta_0\| \xrightarrow{P} 0. \tag{2}$$

The latter reasoning sheds light on the role played by the continuity of the sample paths of $\mathbb{Z}_n$, more specifically, one sees it is indeed at the root of the "no rates" conditions imposed on $\hat{\eta}$. Moreover, conditions (1) and (2) represents a trade-off we need to to accomplish when selecting the norm $\|\cdot\|$. When one prefers to have $\|\cdot\|$ as weak as possible in order to prove (2), one needs the metric to be strong enough so that (1) can hold. In many statistical problems, one is able to show that

$$n^{1/2}(\hat{\theta}(\eta) - \theta_0) = \mathbb{G}_n \varphi_\eta + o_P(1),$$

where the $o_P(1)$ is uniform in η and φ_η is often called the influence function. This asymptotic decomposition of the process $\mathbb{Z}_n$ permit to rely on empirical process theory in order to show (1). As it is summarized by van der Vaart and Wellner (1996), good conditions to impose concerns the metric entropy of the class of functions $\{\varphi_\eta \ : \ \eta \in \mathcal{H}\}$. A relevant paper is van der Vaart and Wellner (2007) in which the authors study conditions implying that $\mathbb{G}_n(\varphi_{\hat{\eta}} - \varphi_{\eta_0})$ goes to 0 in probability.

2 Z-estimators indexed by objective functions

The main purpose of the present work is to establish general conditions for efficiency in a Z-estimation context in which the objective functions are indexed by $\eta \in \mathcal{H}$. More formally, we consider θ_0 and $\hat{\theta}(\eta)$ defined as "zeros" of the maps

$$\theta \mapsto P\psi_\eta(\theta) \qquad \text{and} \qquad \theta \mapsto \mathbb{P}_n\psi_\eta(\theta),$$

where $\psi_\eta(\theta)$ is a measurable map defined on $\mathcal{Z}$. Note that for efficiency purpose η does not affect θ_0. This makes η being a tuning parameter and not a nuisance. Under standard assumptions from the Z-estimation literature (see the article cited bellow), we found that efficiency of $\hat{\theta}(\hat{\eta})$ is guaranteed if $\hat{\eta}$ and $\mathcal{H}$ satisfies Condition (2) and the class

$$\Psi = \{\psi_\eta(\theta) \quad : \quad \theta \in \Theta, \ \eta \in \mathcal{H}\} \quad \text{is } P\text{-Donsker.}$$

Within such a context, characterized by unknown asymptotic distribution, an essential tool to make inference is the bootstrap. The bootstrap empirical process is defined as $\mathbb{P}_n^* f = n^{-1}\sum_{i=1}^n W_{i,n} f(Z_i)$, where the sequence of weights $(W_{i,n})_{i=1,\dots,n}$ is independent from the sequence $(Z_i)_{i=1,\dots,n}$ and satisfies conditions from Praestgaard and Wellner (1993). The bootstrap estimator $\hat{\theta}^*(\eta)$ satisfies

$$\mathbb{P}_n^*\psi_\eta(\theta) = 0.$$

Under mild additional assumptions, we show that the bootstrap works for $\hat{\theta}^*(\hat{\eta})$, i.e. the asymptotic distribution of $\sqrt{n}(\hat{\theta}^*(\hat{\eta}) - \hat{\theta}(\hat{\eta}))$ conditionally on $(Z_i)_{i=1,\dots,n}$, is the same as $\sqrt{n}(\hat{\theta}(\hat{\eta}) - \theta_0)$. Interestingly, a bootstrap of the optimal parameter η_0 is not needed. This common property is due to the asymptotic equicontinuity of the underlying process $\eta \mapsto \mathbb{G}(\psi_\eta(\theta_0))$.

The tools we use in the proof are reminiscent of the following Z-estimation literature. Because Z-estimation theory handles famous statistical methods such as maximum likelihood or least-square estimation, it has received much attention over the last decades. One may first mention the case where θ_0 is Euclidean for which asymptotic normality is obtained for instance in Huber (1987). Later, some authors considered θ_0 an infinite dimensional parameter. Weak convergence with root n rates is obtained in van der Vaart (1995) and the bootstrap is studied in Wellner and Zhan (1996). Taking into account a nuisance parameter with possibly, slower than root n rates of convergence, is developed in Van Keilegom (2003).

3 Weighted regression

The typical applications we have in mind deal with weighted regression. This technique, that attribute different weights to certain observations, is an important tool to handle heteroscedasticity in a data. In the case of a linear regression with $(Y_i, X_i)_{i=1,\dots,n}$, $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^q$, it consists basically in computing

$$\hat{\beta}(w) = \text{argmin}_\beta n^{-1} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2 w(X_i), \tag{3}$$

where w is a real valued function. Among such a collection of estimators, there exists an efficient member $\hat{\theta}(w_0)$ (see Bates and White (1993)). Many studies have focused on the estimation of w_0, for instance, Carroll and Ruppert (1982) argued that a parametric estimation of w_0 can be performed, in Robinson (1987), nonparametric tools are used to approximate w_0. In most of the cases, the authors succeed in showing efficiency. Nevertheless, they rely on relatively long calculations that seems really particular to each context (given by the loss function and the estimator of w_0). Our approach overpass this issue by providing high-level conditions on the estimation of w_0. These conditions are in some ways independent from the rest of the problem. In summary we require that $\|\hat{w} - w_0\| \longrightarrow 0$ in probability (with mild conditions on $\|\cdot\|$) and $P(\hat{w} \in \mathcal{W}) \to 1$ where the class $\mathcal{W}$ is such that, for any $\delta > 0$

$$\int_0^\delta \sup_Q \sqrt{\log \mathcal{N}(\epsilon, \mathcal{W}, L_2(Q))} d\epsilon < +\infty,$$

where the supremum is taken over all finitely discrete probability measures. In a parametric modelling of w_0 the latter conditions are easy to obtain. For nonparametric estimators, one possibility is to ask for smoothness restrictions on the class $\mathcal{W}$ in regards to the dimension q (see Theorem 2.7.1 in van der Vaart and Wellner (1996)).

4 Examples

As it was discussed in the introduction, the results of the paper have applications in showing the efficiency of estimators for which the tuning parameter has been estimated and then plugged-in. This occurs at different levels of statistical theory. We raise in the next several examples.

Example 1 *(Least square constrained estimation)* Let $\hat{\theta}$ be an arbitrary but consistent estimator of θ_0. The estimator $\hat{\theta}_c$ is said to be a least square constraint estimator if it minimizes $(\theta - \hat{\theta})^T \Gamma (\theta - \hat{\theta})$ over $\theta \in \Theta$, where Γ is a symmetric positive definite matrix. Consequently $\hat{\theta}_c$ depends on the choice of Γ but since $|\hat{\theta}_c - \hat{\theta}|_2^2 \leq \text{const.} |\Gamma^{1/2}(\hat{\theta}_c - \hat{\theta})|_2^2 \leq |\Gamma^{1/2}(\theta_0 - \hat{\theta})|_2^2 \to 0$ in probability, the matrix Γ does not change the target quantity θ_0.

It is well know that θ_c is efficient when Γ equals the inverse of the asymptotic variance of $\hat{\theta}$ (see Newey (1994), section 5.2). Such a class is popular among econometricians and known as minimal distance estimator (Newey (1994)).

In the above illustrative example, the use of asymptotic equicontinuity of the process $\Gamma \mapsto \sqrt{n}(\hat{\theta}(\Gamma)-\theta_0)$ is not really legitimate since we can show the efficiency of the procedure using more basic tools such as Slutsky's lemma in Euclidean space. This is due of course to the Euclideanity of θ and Γ but also to the simple form of the map $(\theta, \Gamma) \mapsto (\theta - \hat{\theta})^T \Gamma (\theta - \hat{\theta})$. As a consequence we highlight below more involved examples in which the tuning parameter is either a function (examples 2, 4 and 5) or represents a complicated dependence structure between θ and η (Example 3). To our knowledge, the efficiency property of these examples are quite difficult to obtain.

Example 2 *(weighted regression)* Other losses than the square function can be used and this choice is often related to the noise distribution. Examples include L_p losses, Huber robust loss (see the next example) and least absolute deviation. As in (3), heteroscedasticity can be handled through a weighting of the observations. In a general framework, a formula for the optimal weights is derived in Bates and White (1993).

Example 3 *(Huber cut-off)* Whereas weighted regression handles heteroscedasticity in the data, the cut-off in Huber regression represents the adaptation of the objective function to the distribution of the noise (see Huber (1967)). The Huber objective function is the continuous function that coincides with the identity on $[-c, c]$ (c is called the cut-off) and is constant elsewhere. For instance, a Z-estimator based on this function permits to take care about big tails in the noise distribution by under-weighting large outliers. The choice of the cut-off might be done through an asymptotic variance minimization. An alternative way that seems more "sample-based" consists in minimising the bootstrap approximation of the mean square error.

Example 4 *(instrumental variable)* In Newey (1990), the class of nonlinear instrumental variable is defined through a GMM approach. The estimator $\hat{\theta}$ depends on a so-called "function of instrument" A and satisfies the equation $n^{-1} \sum_{i=1}^{n} A(\widetilde{Z}_i)\psi(Z_i, \theta) = 0$, where each $\widetilde{Z}_i$ a set of coordinates of Z_i. The optimal choice of A can be perform via an analysis of the asymptotic variance.

Example 5 *(dimension reduction)* Sliced inverse regression, introduced by Li (1991), is an important tool for dimension regression. Under the so called linearity condition, the vector $EX\psi(Y)$ describes a subspace of interest when ψ varies. A study of the asymptotic variance leads to an optimal ψ_0. The same can be done for another popular method called sliced average variance estimation.

Bibliographie

[1] Bates, Charles E and White, Halbert (1993), Determination of estimators with minimum asymptotic covariance matrices, Econometric Theory, 9, 04, 633–648.
[2] Carroll, R. J. and Ruppert, D., (1982), Robust estimation in heteroscedastic linear models, The Annals of Statistics, 10, 2, 429–441.
[3] Chen, X. and Linton, O. and Van Keilegom, I. (2003), Estimation of semiparametric models when the criterion function is not smooth, Econometrica, 71, 5, 1591–1608.
[4] Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics, 221–233.
[5] Li, Ker-Chau (1991), Sliced inverse regression for dimension reduction, Journal of the American Statistical Association, 86, 414, 316–342.
[6] Newey, W. K. (1990), Efficient instrumental variables estimation of nonlinear models, Econometrica: Journal of the Econometric Society, 809–837.
[7] Newey, W. K. and McFadden, D. (1994), Large sample estimation and hypothesis testing, Handbook of econometrics, Vol. IV, 2, 2111–2245.
[8] Præstgaard, J. and Wellner, J. A. (1993), Exchangeably weighted bootstraps of the general empirical process, The Annals of Probability, 2053–2086.
[9] Robinson, P. M. (1987), Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form, Econometrica, 55, 4, 875–891.
[10] van der Vaart, A. W. (1995), Efficiency of infinite-dimensional M-estimators, Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research, 49, 1, 9–30.
[11] van der Vaart, A. W. and Wellner, J. A. (1996), Weak convergence and empirical processes, Springer Series in Statistics, Springer-Verlag.
[12] van der Vaart, A. W. and Wellner, J. A. (2007), Empirical processes indexed by estimated functions, Asymptotics: particles, processes and inverse problems, 55, 234–252.
[13] Wellner, J. A. and Zhan, Y. (1996), Bootstrapping Z-estimators, University of Washington Department of Statistics Technical Report, 308.