Contents

2 September 2013

Chapter 1

M-estimation

1.1 From model to estimator

Mestimation::Mest

This handout describes a general way of constructing pretty good estimators for models where $\mathcal{X} = \mathbb{R}^n$ with generic point $x = (x_1, \ldots, x_n)$ and where, under each $\mathbb{P}_\theta$, the $x_1, \ldots, x_n$ are independent, each distributed according to some probability distribution P_θ on $\mathbb{R}$.

Note well that P_θ is a probability distribution on the real line and $\mathbb{P}_\theta$ is a joint distribution. For example, if P_θ is specified by a density $p(z, \theta)$ on the real line then $\mathbb{P}_\theta$ is specified by the joint density

$$p(x_1, \theta)p(x_2, \theta) \ldots p(x_n, \theta).$$

I assume that the index set Θ for the model is a subset of some Euclidean space, $\mathbb{R}^L$, and that $\tau : \Theta \to T \subseteq \mathbb{R}^k$ is the quantity that is to be estimated. (In class I will talk mostly about the case $k = L = 1$.)

Remark. Each x_i takes values in the real line, $\mathbb{R}$. I could replace this $\mathbb{R}$ by some other Euclidean space, or something even fancier, without much effect on the argument that follows.

Some authors write $\mathbb{P}_{\theta,n}$ instead of just $\mathbb{P}_\theta$, to stress that the theory concerns a sequence of models based on increasing sample sizes.

Consider a function $g : \mathbb{R} \times T \to \mathbb{R}$. Define

$$G(\theta, t) = \mathbb{E}_\theta g(x_1, t).$$

In this handout I consider the case where g can be chosen so that $t \mapsto G(\theta, t)$ is minimized at $t = \tau(\theta)$, for each θ. That is,

tau.def <1>

$$\tau(\theta) = \operatorname*{argmin}_{t \in T} G(\theta, t) \qquad \text{for each } \theta.$$

If you prefer, you could actually define τ by the last equality.

To estimate $\tau(\theta)$ first define

$$G_n(t) = G_n(t, x_1, \ldots, x_n) := n^{-1} \sum\nolimits_{i \le n} g(x_i, t).$$

Suppose G_n is minimized at some point $\widehat{\tau}_n$ of T, that is,

$$\widehat{\tau}_n(x_1, \ldots, x_n) = \operatorname*{argmin}_{t \in T} G_n(t, x_1, \ldots, x_n)$$

The random variable $\widehat{\tau}_n(x_1, \ldots, x_n)$ is called an ***M-estimator*** because it is defined by a minimization operation. Some authors prefer to deal with maximization operations. The two approaches are equivalent: just replace g by $-g$.

In this handout I will show, under broad assumptions, when sampling from the P_θ distribution the $\widehat{\tau}_n$ is close to $\tau(\theta)$ with high $\mathbb{P}_\theta$ probability if n is large enough.

The idea is that, in the limit as the sample size goes to infinity, we learn the whole function $G(\theta, .)$ excatly if we are sampling from P_θ. A simple minimization would then give us the value $\tau(\theta)$. If we know $G(\theta, \cdot)$ only approximately (based on a finite sample size) then we can determine $\tau(\theta)$ only approximately.

median <2> **Example.** For z and t in $\mathbb{R}$ define $g(z,t) = |z-t|-|z|$. If $G(t) = \mathbb{E}_P g(z,t) = \int g(z,t) dP$, show that G is minimized at a median of P. See Homework 1.

Remark. Here and elsewhere you should interpret $\int g(z,t)dP$ as $\sum_j g(z_j,t)P\{z_j\}$ if P is a discrete distribution with atoms at $z_1, z_2, \ldots$ and as $\int g(z,t)p(z)\,dz$ if P is a continuous distribution with density p.

That is, the $\tau(\theta)$ defined by equality <1> is equal to the median of the P_θ distribution. Implicitly, this definition assumes that the median is unique for the "population" P_θ.

For the sample the median is not unique when n is even. I intend to ignore that small detail.

□

mle.pop <3> **Example.** Consider the case where $T = \Theta$ for some subset Θ of $\mathbb{R}^L$. Suppose $p(z, \theta)$ is a probability density for each θ in Θ. Define $g(z,t) =$

$-\ell(z,t)$ where $\ell(z,\theta) = \log p(z,\theta)$. Write $\dot{p}(z,\theta)$ for $\partial p(z,\theta)/\partial\theta$ and $\dot{\ell}(z,\theta)$ for $\partial\ell(z,\theta)/\partial\theta = \dot{p}(z,\theta)/p(z,\theta)$. Then

$$\frac{\partial G(\theta,t)}{\partial t} = -\int p(z,\theta)\frac{\partial}{\partial t}\ell(z,t) = \int p(z,\theta)\frac{\dot{p}(z,t)}{p(z,t)}.$$

If we put t equal to θ then factors cancel, leaving

$$\frac{\partial G(\theta,t)}{\partial t}\Big|_{t=\theta} = \int \frac{\partial}{\partial t}p(z,t)\Big|_{t=\theta} = \left(\frac{\partial}{\partial t}\int p(z,t)\right)\Big|_{t=\theta} = 0.$$

The last equality comes from the fact that $\int p(z,t) = 1$ for every t.

That is, when $g = -\ell$ we have $\tau(\theta) = \theta$. The estimator $\widehat{\theta}_n$ minimizes

$$n^{-1}\sum\nolimits_{i\le n} \log p(x_i,t) \qquad \text{over all } t \text{ in } \Theta.$$

Equivalently, it maximizes the joint density

$$p(x_1,t)\ldots p(x_n,t) \qquad \text{over all } t \text{ in } \Theta.$$

That is, $\widehat{\theta}_n$ is the ***maximum likelihood estimator*** (MLE)

□

Remark. The quantity

$$G(\theta,t) - G(\theta,\theta) = \int p(z,\theta)\log\left(p(z,\theta)/p(z,t)\right)\,dz$$

is often called the Kullback-Leibler distance (or relative entropy) between P_θ and P_t, often denoted by $K(P_\theta, P_t)$. It can be shown using Jensen's inequality that $K(P,Q)$, for probability distributions P and Q, is always nonnegative. If P and Q are given by densities p and q, then it can also be shown that

$$K(P,Q) \ge \tfrac{1}{2}\left(\int |p-q|\right)^2,$$

a result known as Pinsker's inequality.

1.2 Probability facts

Mestimation::probfacts

The asymptotic theory for M-estimators depends mostly on three important probability tools. The fourth result (Jensen's inequality) in this Section is there just for future reference.

LLN

1.2.1 Law of Large Numbers

For independent random variables $Y_1, Y_2, \ldots$ each with the same distribution, the Law of Large Numbers (LLN) asserts that, in various probabilistic senses

$$n^{-1} \sum\nolimits_{i \le n} Y_i \to \mathbb{E} Y_1 \qquad \text{as } n \to \infty.$$

CLT

1.2.2 Central Limit Theorem

If $\mathbb{E} Y_1 = 0$ and $\sigma^2 = \operatorname{var}(Y) < \infty$ then the Central Limit Theorem (CLT) asserts that

$$n^{-1/2} \sum\nolimits_{i \le n} Y_i \dot{\sim} N(0, \sigma^2),$$

in the sense that the distribution approaches normality as n gets larger. Similar assertions hold for random vectors. For example, if the Y_i's are identically distributed random vectors with zero expected value and variance matrix $V = \mathbb{E}(Y_1 Y_1')$ then

$$n^{-1/2} \sum\nolimits_{i \le n} Y_i \dot{\sim} N(0, V),$$

with $N(0, V)$ denoting a multivariate normal with zero mean and variance matrix V.

CS

1.2.3 Cauchy-Schwarz inequality

If X and Y are random variables then

$$|\mathbb{E}(XY)| \le \sqrt{\mathbb{E}(X^2)} \sqrt{\mathbb{E}(Y^2)}$$

Here is a quick proof, just in case you haven't seen the inequality before.

Write a for $\sqrt{\mathbb{E}(X^2)}$ and b for $\sqrt{\mathbb{E}(Y^2)}$. Then

$$0 \le \mathbb{E} \left(\frac{X}{a} \pm \frac{Y}{b} \right)^2 = 1 \pm 2\mathbb{E}(XY)/(ab) + 1,$$

which rearranges to $2ab \ge \mp 2\mathbb{E}(XY)$.

Notice that the Cauchy-Scwarz inequality becomes an equality if and only if either $bX = aY$ or $bX = -aY$.

Jensen

1.2.4 Jensen's inequality

If Ψ is a convex function then

$$\mathbb{E}\Psi(X) \geq \Psi(\mathbb{E}X).$$

For the special case where $\psi(z) = z^2$ the inequality becomes $\mathbb{E}(X^2) \geq (\mathbb{E}X)^2$, which is equivalent to the fact that $\operatorname{var}(X) \geq 0$.

Mestimation::consistency

1.3 Consistency

The Law of Large Numbers for each fixed t gives $G_n(t) \to G(\theta, t)$ as $n \to \infty$ under $\mathbb{P}_\theta$. Something a little more than pointwise convergence gives

$$\tau_n = \operatorname*{argmin}_{t\in T} G_n(t) \to \operatorname*{argmin}_{t\in T} G(\theta, t) = \tau(\theta)$$

in some probabilistic sense. For example, $\{\widehat{\tau}_n\}$ is said to be ***(weakly) consistent*** for $\tau(\theta)$ if

$$\mathbb{P}_\theta\{|\widehat{\tau}_n - \tau(\theta)| > \epsilon\} \to 0 \qquad \text{for each } \epsilon > 0 \text{ and each } \theta \in \Theta.$$

Mestimation::asynorm

1.4 Asymptotic normality

Once we know that $\widehat{\theta}$ concentrates near $\tau(\theta)$ it makes sense to look at approximations to $G_n(t)$ for t in a neighborhood of $\tau(\theta)$.

The analysis in this subsection is carried out for a fixed θ. To avoid the temptation to differentiate with respect to a fixed θ, I will temporariliy omit it from the notation, writing P instead of P_θ and $G(t)$ instead of $G(\theta, t)$ and τ instead of $\tau(\theta)$.

Remark. In fact the analysis works even if P is not one of the distributions specified by the model. It matters only that the function $G(t) = \mathbb{E}g(x_1, t) = \int g(x_1, t)\, dP$ has a unique maximum at τ.

Assuming that g is a smooth function of t, make a Taylor expansion of g around τ.

Taylor <4> $$g(x_1,t) \approx g(x_1,\tau) + (t-\tau)'\dot{g}(x_1,\tau) + \tfrac{1}{2}(t-\tau)'\ddot{g}(x_1,\tau)(t-\tau) \qquad \text{for } t \text{ near } \tau.$$

Here $\dot{g}(x_1, \tau)$ denotes the $k\times 1$ vector of functions $\partial g(x_1, t)/\partial t_i$, for $1 \leq i \leq k$, and $\ddot{g}(x_1, \tau)$ denotes the $k \times k$ matrix of functions $\partial^2 g(x_1, t)/\partial t_i \partial t_j$, for $1 \leq i, j \leq k$, and $'$ denotes transpose.

Remark. I am using the $\bullet$ to denote differentiation instead of $'$, which gets confused with transpose. If you are not used to Taylor expansions of vector-valued functions you could just assume $k = 1$.

If we replace x_1 by x_i then average over $1 \le i \le n$ approximation <4> gives

Gn.approx <5>
$$G_n(t) \approx G_n(\tau) + (t-\tau)n^{-1/2}Z_n + \tfrac{1}{2}(t-\tau)'J_n(t-\tau)$$

where

$$Z_n = n^{-1/2}\sum\nolimits_{i\le n} \dot{g}(x_i,\tau) \quad \text{AND} \quad J_n = n^{-1}\sum\nolimits_{i\le n} \ddot{g}(x_i,\tau).$$

You'll see soon why Z_n only needs the $n^{-1/2}$ rescaling.

If we take expectations of both sides of <4>, ignoring the remainder terms we get a Taylor expansion of G around τ,

$$G(t) \approx G(\tau) + (t-\tau)'\mathbb{E}\dot{g}(x_1,\tau) + \tfrac{1}{2}(t-\tau)'\mathbb{E}\ddot{g}(x_1,\tau)(t-\tau)$$

Compare with

$$G(t) \approx G(\tau) + \tfrac{1}{2}(t-\tau)'\ddot{G}(\tau)(t-\tau).$$

The linear term, $(t-\tau)'\dot{G}(\tau)$, vanishes because τ is the minimimizing value for G.

Remark. Here I am assuming that τ lies in the interior of T, so that minimization implies a zero derivative. Very strange things can happen when τ lies on the boundary of T

Deduce that

zero.mean <6>
$$\mathbb{E}\dot{g}(x_1,\tau) = \dot{G}(\tau) = 0.$$

Similarly, the $k \times k$ matrix

$$J = \mathbb{E}\ddot{g}(x_1,\tau) = \ddot{G}(\tau)$$

must be at least nonnegative definite, that is

$$\delta' J \delta \ge 0 \qquad \text{for all } \delta \in \mathbb{R}^k,$$

otherwise there would be some direction along which G decreased below its minimum.

Remark. If there is some $\delta \neq 0$ for which $\delta' J \delta = 0$ the theory gets more complicated. (The matrix J is then singular and does not have an inverse.) Textbooks seldom mention that possibility. I'll ignore it too. That is, I'll assume J is actually positive definite with $\delta' J \delta > 0$ for all nonzero δ.

The LLN (see subsection 1.2.1) applied to each of the k^2 entries of the $k \times k$ matrix J_n from <5> gives $J_n \approx J$, with an error of approximation that goes to zero in some suitable probabilistic sense. The Z_n is an average of random vectors with zero expected values, by <6>, so the CLT gives

Zn <7> $$Z_n \dot{\sim} N(0, V) \qquad \text{with } V = \mathbb{E}\dot{g}(x_1, \tau)\dot{g}(x_1, \tau)'.$$

Approximation <5> can be rewritten as

$$G_n(t) \approx G_n(\tau) + (t-\tau)' n^{-1/2} Z_n + \tfrac{1}{2}(t-\tau)' J (t-\tau) \qquad \text{for } t \text{ near } \tau.$$

The quadratic on the right-hand side can be simplified by writing s for $t - \tau$ and W for $n^{-1/2} Z_n$, leaving

$$\tfrac{1}{2} s' J s + s' W = \tfrac{1}{2}(s + J^{-1}W)' J (s + J^{-1}W) - \tfrac{1}{2} W' J^{-1} W$$

on the right-hand side. Positive definiteness of J ensures that the quadratic is minimized when $s = -J^{-1}W$. Assuming that the approximation of G_n translates into approximation of its minimizer, conclude that $\widehat{\tau}_n - \tau \approx -n^{-1/2} J^{-1} Z_n$, so that

limit.distn <8> $$n^{1/2}(\widehat{\tau}_n - \tau) \approx -J^{-1} Z_n \dot{\sim} N(0, J^{-1} V J^{-1})$$

with

$$J = \mathbb{E}\ddot{g}(x_1, \tau) \quad \text{AND} \quad V = \mathbb{E}\dot{g}(x_1, \tau)\dot{g}(x_1, \tau)'.$$

mle.sample <9> **Example.** Consider once more the maximum likelihood estimator from Example <3>, where $T = \Theta \subseteq \mathbb{R}^L$ and $g(z,t) = -\ell(z,t) = -\log p(z,t)$. Remember that $\tau(\theta) = \theta$ and that I was writing $\widehat{\theta}_n$ instead of $\widehat{\tau}_n$. Remember also (cf. equality <6>) that $\mathbb{E}_\theta \dot{\ell}(x_1, \theta) = 0$.

For this special case, the $L \times L$ variance matrix $V(\theta)$ equals

$$\mathbb{E}\dot{g}(x_1,\tau)\dot{g}(x_1,\tau)' = \operatorname{var}\left(\dot{\ell}(x_1,\tau)\right) =: \mathbb{I}(\theta).$$

The matrix $\mathbb{I}(\theta)$ is called the ***Fisher information matrix***.

The $J(\theta) = \mathbb{E}_\theta \ddot{\ell}(x_1,\theta)$ also takes a special form. First note that

$$\ddot{\ell}(z,t) = \frac{\partial \dot{\ell}(z,t)}{\partial t} = \frac{\partial}{\partial t}\frac{\dot{p}(z,t)}{p(z,t)} = \frac{\ddot{p}(z,t)}{p(z,t)} - \dot{p}(z,t)\dot{p}(z,t)'/p(z,t)^2.$$

Take expectations.

$$\mathbb{E}_\theta \ddot{\ell}(x_1,\theta) = \int \partial^2 p(z,t)/\partial t^2\big|_{t=\theta} - \int p(z,\theta)\dot{\ell}(z,\theta)\dot{\ell}(z,\theta)'.$$

The first term vanishes—takes the second derivative outside the integral sign then use the fact that $\int p(z,t) = 1$ for all t. The second term is just the information matrix again. In summary, $V(\theta) = -J(\theta) = \mathbb{I}(\theta)$. Approximation <8> becomes

MLE.limit.distn <10>

$$\text{MLE:} \qquad Z_n = -n^{-1/2}\textstyle\sum_{i\le n}\dot{\ell}(x_i,\theta) \dot{\sim} N(0,\mathbb{I}(\theta)) \text{ and}$$

$$n^{1/2}\left(\widehat{\theta}_n - \theta\right) \approx -\mathbb{I}(\theta)^{-1}Z_n \dot{\sim} N(0,\mathbb{I}(\theta)^{-1}) \qquad \text{under } \mathbb{P}_\theta,$$

a very famous approximation.

□

1.5 Asymptotic efficiency

Mestimation::efficiency

Classical (Fisherian) statistical theory assigns the MLE a very special role. According to Fisher, the MLE minimizes the limiting variance amongst all estimators. He called this property (asymptotic) ***efficiency***. Modern theory has shown Fisher's assertion to be wrong unless hedged with further restrictions. One such restriction is to consider only M-estimators as competitors.

Remark. Actually Fisher usually didn't include the word "asymptotic". I find that this omission leads to a lot of confusion between optimality properties for fixed, finite sample size and optimality properties for the limiting distribution.

For this section suppose P_θ is specified by some density function $p(z,\theta)$ on the real line. (The argument for discrete distributions is similar.)

Once again consider the problem of estimating $\tau(\theta) = (\tau_1(\theta), \ldots, \tau_k(\theta))$, based on independent samples $x_1, x_2, \ldots$ from P_θ. Assume τ is a smooth function of θ, so that the $k \times L$ matrix $\dot{\tau}(\theta)$ with $\tau_{ij}(\theta) = \partial\tau_i/\partial\theta_j$ is well defined.

For each fixed δ in $\mathbb{R}^k$, approximation <8> implies

$$n^{1/2}\delta'\left(\widehat{\tau}_n - \tau(\theta)\right) \dot{\sim} N(0, \sigma^2(\theta)) \qquad \text{under } \mathbb{P}_\theta$$

where

$$\begin{aligned}
\sigma^2_\delta(\theta) &:= \delta' J(\theta)^{-1} V(\theta) J(\theta)^{-1}\delta \\
J(\theta) &= \mathbb{E}_\theta \ddot{g}(x_1, \tau(\theta)) \\
V(\theta) &= \mathbb{E}\dot{g}(x_1,\tau)\dot{g}(x_1,\tau(\theta))'.
\end{aligned}$$

The aim is to find a g to minimize the asymptotic variance $\sigma^2_\delta(\theta)$.

The first step is to find a lower bound for $\sigma^2(\theta)$ by considering the behaviour of $\widehat{\tau}_n$ for values of θ of the form

$$\theta_s = \theta + s\gamma \qquad \text{for } s \in \mathbb{R}.$$

Here θ denotes some point of Θ that is fixed throughout the argument and γ is a nonzero vector in $\mathbb{R}^L$. Notice that

$$\frac{\partial}{\partial s}\tau(\theta_s) = \dot{\tau}(\theta_s)\gamma,$$

the product of a $k \times L$ matrix with an $L \times 1$ vector.

Along the path defined by θ_s inequality <6> becomes

$$0 = \mathbb{E}_{\theta_s}\dot{g}(x_1, \tau(\theta_s)) = \int \dot{g}(x_1, \tau(\theta_s))p(z, \theta_s) \qquad \text{for all } s.$$

Notice the way θ_s appears in two places. Differentiate with respect to s.

$$\begin{aligned}
0 &= \int \ddot{g}(x_1, \tau(\theta_s))\dot{\tau}(\theta_s)\gamma\, p(z,\theta_s) + \int \dot{g}(x_1,\tau(\theta_s))\dot{p}(z,\theta_s)'\gamma \\
&= \mathbb{E}_{\theta_s}\left(\ddot{g}(x_1,\tau(\theta_s))\right)\dot{\tau}(\theta_s)\gamma + \mathbb{E}_{\theta_s}\left(\dot{g}(x_1,\tau(\theta_s))\dot{\ell}(z,\theta_s)'\gamma\right)
\end{aligned}$$

Put $s = 0$ then multiply both sides by $\delta' J(\theta)^{-1}$ to deduce

constrain <11>
$$\delta'\dot{\tau}(\theta)\gamma = -\mathbb{E}_\theta\left(\delta' J(\theta)^{-1}\dot{g}(x_1,\tau(\theta))\dot{\ell}(z,\theta)'\gamma\right).$$

The right-hand side is of the form $\mathbb{E}_\theta(XY)$ for random variables

$$X = \delta' J(\theta)^{-1} \overset{\bullet}{g}(x_1, \tau(\theta)) \quad \text{AND} \quad Y = \overset{\bullet}{\ell}(z, \theta)'\gamma.$$

Notice that

$$\mathbb{E}_\theta X^2 = \sigma_\delta^2(\theta) \quad \text{AND} \quad \mathbb{E}_\theta Y^2 = \gamma' \mathbb{I}(\theta)\gamma.$$

Invoke the Cauchy-Schwarz inequality (subsection 1.2.3) to deduce that

$$\left(\delta' \overset{\bullet}{\tau}(\theta)\gamma\right)^2 \le \sigma_\delta^2(\theta)\, \gamma' \mathbb{I}(\theta)\gamma$$

That is,

$$\sigma_\delta^2(\theta) \ge \frac{\left(\delta' \overset{\bullet}{\tau}(\theta)\gamma\right)^2}{\gamma' \mathbb{I}(\theta)\gamma} \qquad \text{for all } \gamma \ne 0.$$

Replace γ by $\mathbb{I}(\theta)^{-1/2}\beta$ then choose β as the unit vector in the direction $\delta' \overset{\bullet}{\tau}(\theta)\mathbb{I}(\theta)^{-1/2}$ to maximize the right-hand side, leaving the lower bound

var.lower <12>
$$\delta' J(\theta)^{-1} V(\theta) J(\theta)^{-1}\delta = \sigma_\delta^2(\theta) \ge \lambda_\delta(\theta) := \delta' \overset{\bullet}{\tau}(\theta) \mathbb{I}^{-1}(\theta) \overset{\bullet}{\tau}(\theta)'\delta.$$

1.5.1 Special case: $\tau(\theta) = \theta$

When $\tau(\theta) = \theta$ the matrix $\overset{\bullet}{\tau}(\theta)$ of derivatives becomes the identity matrix I_L and the right-hand side of <12> becomes $\delta' \mathbb{I}(\theta)^{-1}\delta$. For the special case of the MLE $\widehat{\theta}_n$ the left-hand side of <12> also equals $\delta' \mathbb{I}(\theta)^{-1}\delta$. That is, the MLE achieves the lower bound, *for every* δ.

1.5.2 General case

When $\overset{\bullet}{\tau}(\theta)$ is not the identity matrix it is not obvious to me how to find an M-estimator that achieves the lower bound in <12>. But there is another way to estimate $\tau(\theta)$ that does give an asymptotic variance equal to $\lambda_\delta(\theta)$.

By Taylor's Theorem,

$$\tau(\theta + h) \approx \tau(\theta) + \overset{\bullet}{\tau}(\theta)h \qquad \text{for small } h \in \mathbb{R}^L.$$

In particular, for $h = \widehat{\theta}_n - \theta$ the approximation <10> for the MLE $\widehat{\theta}_n$ gives

$$\tau(\widehat{\theta}_n) - \tau(\theta) \approx -\dot{\tau}(\theta) n^{-1/2} \mathbb{I}(\theta)^{-1} Z_n$$

where $Z_n = -n^{-1/2} \sum_{i \le n} \dot{\ell}(x_i, \theta) \dot{\sim} N(0, \mathbb{I}(\theta))$. Thus

$$n^{1/2} \left(\tau(\widehat{\theta}_n) - \tau(\theta) \right) \approx \dot{\tau}(\theta) \mathbb{I}(\theta)^{-1} Z_n \dot{\sim} N(0, W(\theta)) \qquad \text{under } \mathbb{P}_\theta,$$

where

$$W(\theta) = \dot{\tau}(\theta) \mathbb{I}(\theta)^{-1} \mathbb{I}(\theta) \mathbb{I}(\theta)^{-1} \dot{\tau}(\theta)'.$$

In consequence, $n^{1/2} \delta' \left(\tau(\widehat{\theta}_n) - \tau(\theta) \right) \dot{\sim} N(0, \lambda_\delta(\theta))$.