Estimation process, consistency

Petr Lachout
lachout@karlin.mff.cuni.cz
Charles University in Prague

Tartu 2007

Introduction

We consider a general scheme of parameter estimation.
Our task is to estimate true value of a parameter.

Let us denote it θ_0. We suppose to know the set, say Θ, of all possible values of this parameter and a parameterized family of probability measures $\mathcal{P}_\Theta = \{\mu_\theta \mid \theta \in \Theta\}$ defined on a metric space $\mathcal{Y}$.
In any time $t \in \mathbb{N}$ we have known an observed data $Z_t \in \mathcal{Z}_t$.

Typically, we observe a sequence of data $X_1, X_2, X_3, \ldots$.
belonging to a metric space $\mathcal{X}$.
We group observations available at time $t \in \mathbb{N}$ in a vector $Z_t = (X_1, X_2, \ldots, X_{k_t})$ and $\mathcal{Z}_t = \mathcal{X}^{k_t}$.

From observed data we construct probability measure $\mu_t(\bullet \mid Z_t)$ on $\mathcal{Y}$. These measures will play role of estimators for the "true" probability measure μ_{θ_0}.
The true parameter θ_0 is estimated by an ε_t-estimator $\hat{\theta}_t \in \Theta$, i.e. fulfilling for all $\theta \in \Theta$

$$\mathsf{L}(\mu_t(\bullet \mid Z_t)\,;\hat{\theta}_t) < \mathsf{L}(\mu_t(\bullet \mid Z_t)\,;\theta) + \varepsilon_t, \tag{1}$$

where L is a given "distance" between measures and parameters.

For our purposes we need a bit stronger notion of the standard weak convergence of probability measures. Therefore we have to introduce a convenient notation.

Definition

Let μ, μ_n, $n \in \mathbb{N}$ be Borel probability measures on a metric space $\mathcal{Y}$ and $\mathcal{F} \subset \{f : \mathcal{Y} \to \mathbb{R} \mid f \text{ is measurable}\}$. We will say that μ_n converge $\mathcal{F}$-weakly to μ iff

$$\int_{\mathcal{Y}} f(y)\mu_n(\mathrm{d}y) \xrightarrow[n\to+\infty]{} \int_{\mathcal{Y}} f(y)\mu(\mathrm{d}y) \qquad \text{for all bounded continuous function } f : \mathcal{Y} \to \mathbb{R};$$

$$\int_{\mathcal{Y}} f(y)\mu_n(\mathrm{d}y) \xrightarrow[n\to+\infty]{} \int_{\mathcal{Y}} f(y)\mu(\mathrm{d}y) \qquad \text{for all } f \in \mathcal{F}.$$

We will denote the convergence by

$$\mu_n \xrightarrow[n\to+\infty]{\mathcal{F}-w} \mu.$$

For example, the strong law of large numbers for i.i.d. real random variables can be rewritten as $\nu_n \xrightarrow[n\to+\infty]{\mathcal{H}-w} \nu$, where ν_n is the empirical measure defined from observations till time n, ν is common distribution of the observations and $\mathcal{H} = \{h : x \in \mathbb{R} \to |x|\}$.

Consequently, we have

$$\int_{\mathbb{R}} f(y)\nu_n(\mathrm{d}y) \xrightarrow[n\to+\infty]{} \int_{\mathbb{R}} f(y)\nu(\mathrm{d}y)$$

for every continuous function $f : \mathbb{R} \to \mathbb{R}$ fulfilling $|f(y)| \leq A + B|y|$ for all $y \in \mathbb{R}$ and convenient $A, B \in \mathbb{R}$.

General result

Now, let us formalize the schema in a list of assumptions.

Assumption A1

Spaces $\mathcal{Y}$, Φ, $\mathcal{Z}_t$, $t \in \mathbb{N}$ are metric spaces. The set $\Theta \subset \Phi$ is nonempty and $\mathcal{F} \subset \{f : \mathcal{Y} \to \mathbb{R} \mid f \text{ is measurable}\}$.
(The set $\mathcal{F}$ is allowed to be empty.)

Assumption A2

$\varepsilon_t : \Omega \to \mathbb{R}_{++}$ for any $t \in \mathbb{N}$ and $\bar{\varepsilon} = \limsup_{t \to +\infty} \varepsilon_t < +\infty$.

Assumption A3

For any $\theta \in \Theta$, μ_θ is a Borel probability measure on $\mathcal{Y}$.

Assumption A4

For any $t \in \mathbb{N}$, we observe $Z_t : \Omega \to \mathcal{Z}_t$.

Assumption A5

For any $t \in \mathbb{N}$, $z_t \in \mathcal{Z}_t$, $\mu_t(\bullet \,|\, z_t)$ is a Borel probability measure on $\mathcal{Y}$.

We denote $\mathcal{P}_{emp} = \{\mu_t(\bullet \,|\, z_t) \mid z_t \in \mathcal{Z}_t, t \in \mathbb{N}\}$.

Assumption A6

The function $\mathrm{L} \ : \ (\mathcal{P}_{emp} \cup \mathcal{P}_\Theta) \times \Theta \to \mathbb{R}$ is non-negative.

Assumption A7

θ_0 is a minimizer of the function $\mathrm{L}(\mu_{\theta_0}\,; \bullet)$. In other words $\theta_0 \in \operatorname{argmin} \{\mathrm{L}(\mu_{\theta_0}\,; \theta) \mid \theta \in \Theta\}$.

Assumption A8

Whenever $\forall n \in \mathbb{N}\ \nu_n \in \mathcal{P}_{emp}$ and $\nu_n \xrightarrow[n\to+\infty]{\mathcal{F}-w} \mu_{\theta_0}$, then there is a sequence $\tilde{\theta}_n \in \Theta$, $n \in \mathbb{N}$ such that

$$\lim_{n\to+\infty} \tilde{\theta}_n = \theta_0 \quad \text{and} \quad \lim_{n\to+\infty} \mathsf{L}(\nu_n\,;\tilde{\theta}_n) = \mathsf{L}(\mu_{\theta_0}\,;\theta_0).$$

Assumption A9

There is a compact set $K \subset \Theta$ such that

1. $\liminf\limits_{n\to+\infty} \mathrm{L}(\nu_n\,;\theta_n) \geq \mathrm{L}(\mu_{\theta_0}\,;\theta)$ whenever
   $\forall n \in \mathbb{N}\ \nu_n \in \mathcal{P}_{emp}$ and $\nu_n \xrightarrow[n\to+\infty]{\mathcal{F}-w} \mu_{\theta_0}$,
   $\forall n \in \mathbb{N}\ \theta_n \in \Theta$, $\theta_n \xrightarrow[n\to+\infty]{} \theta \in K$.
2. For any sequence of probability measures $\nu_n \in \mathcal{P}_{emp}$,
   $\nu_n \xrightarrow[n\to+\infty]{\mathcal{F}-w} \mu_{\theta_0}$ and any open set $G \supset K$ we have

$$\liminf_{n\to+\infty} \inf_{\theta\in\Theta\setminus G} \mathrm{L}(\nu_n\,;\theta) > \mathrm{L}(\mu_{\theta_0}\,;\theta_0) + \bar{\varepsilon}.$$

These assumptions ensures existence of the estimator and also that it is a consistent estimator of the true parameter.

Lemma

Let Assumptions A1–A6 be fulfilled. Then an ε_t-estimator $\hat{\theta}_t$ fulfilling (1) exists for any $t \in \mathbb{N}$.

Proof.

Let $t \in \mathbb{N}$. Accordingly to Assumptions A6 and A2,

$$0 \leq \inf_{\theta \in \Theta} \mathsf{L}(\mu_t(\bullet \mid Z_t)\,;\theta) < +\infty \quad \text{and} \quad \varepsilon_t > 0.$$

Hence, a $\hat{\theta}_t$ fulfilling (1) always exists. $\square$

We have to recall a few from topological terminology.

Definition

For a sequence η_n, $n \in \mathbb{N}$ in a metric space $\mathcal{W}$, we denote the set of its cluster points by $\mathrm{Ls}\left(\eta_n\,,\; n \in \mathbb{N}\right)$, *i.e.*

$$\mathrm{Ls}\left(\eta_n\,,\; n \in \mathbb{N}\right) = \left\{ \psi \in \mathcal{W} \;\middle|\; \exists \textit{ subsequence s.t. } \lim_{n \to +\infty} \eta_{k_n} = \psi \right\}.$$

Definition

We say that a sequence η_n, $t \in \mathbb{N}$ in a metric space $\mathcal{W}$ is compact if each its subsequence possesses at least one cluster point.

Compact sequence in metric space possesses an equivalent description.

Lemma

Let η_n, $t \in \mathbb{N}$ be a sequence in a metric space $\mathcal{W}$. Then, the following statements are equivalent:

1. *The sequence is compact.*
2. *There is a compact $L \subset \mathcal{W}$ such that $\eta_n \in L$ for all $t \in \mathbb{N}$.*
3. *The set $\{\eta_n \mid n \in \mathbb{N}\} \cup \mathrm{Ls}\left(\eta_n\,,\ n \in \mathbb{N}\right)$ is compact.*

Lemma

Let η_n, $n \in \mathbb{N}$ *be a sequence in a metric space* $\mathcal{W}$ *and* $K \subset \mathcal{W}$ *be a compact. If for every open set* $G \supset K$ *there is an* $n_G \in \mathbb{N}$ *such that* $\eta_n \in G$ *for all* $n \in \mathbb{N}$, $n \geq n_G$.
Then the sequence is compact and $\mathrm{Ls}\left(\eta_n\,,\ n \in \mathbb{N}\right) \subset K$.

Theorem

Let $\Omega_0 \subset \Omega$ be such that for all $\omega \in \Omega_0$

$$\mu_t(\bullet \,|\, Z_t(\omega)) \xrightarrow[n\to+\infty]{\mathcal{F}-w} \mu_{\theta_0}$$

and Assumptions A1-A9 be fulfilled.
Then $\theta_0 \in K$ and $\hat{\theta}_t$ exists for any $t \in \mathbb{N}$. Further, for all $\omega \in \Omega_0$ the sequence $\hat{\theta}_t(\omega)$, $t \in \mathbb{N}$ is compact and

$$\emptyset \neq \mathsf{Ls}\left(\hat{\theta}_t(\omega)\,,\; t \in \mathbb{N}\right) \subset \left\{\theta \in K \;\middle|\; \mathsf{L}(\mu_{\theta_0}\,;\theta) \leq \mathsf{L}(\mu_{\theta_0}\,;\theta_0) + \bar{\varepsilon}(\omega)\right\}.$$

Our proof treats any trajectory separately. Therefore, we do not need measurability of $\mu_t(\bullet \mid z_t)$ with respect to $z_t \in \mathcal{Z}_t$. Also, the definition of the ε_t-estimator does not require measurability. Thus, it can naturally happen that the estimator is not a random variable.

Linear regression

We suppose to observe couples $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_t, X_t)$ connected by a linear regression model

$$Y_i = X_i^\top \beta_0 + \mathrm{e}_i \quad \forall\, i = 1, 2, \ldots, t. \tag{3}$$

Where $Y_i : \Omega \to \mathbb{R}$, $X_i : \Omega \to \mathbb{R}^d$ are mappings, $\mathrm{e}_i : \Omega \to \mathbb{R}$ are unobserved mappings and $\beta_0 \in \Theta \subset \mathbb{R}^d$ is deterministic but unknown parameter.

As probability measures required in Assumption A5 we will employ empirical probability measure defined from observations. Let us define denotation of an empirical probability measure in a general case. Let $\mathcal{W} \neq \emptyset$ and $w_1, w_2, \ldots, w_t \in \mathcal{W}$. Then the empirical probability measure is defined for any $A \subset \mathcal{W}$ as the relative number of observations hitting the set A, i.e. by the formula

$$\mathcal{E}_t(A \mid w_1, w_2, \ldots, w_t) = \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}[w_i \in A]. \tag{4}$$

Let us recall that if $\mathcal{W}$ is a metric space then empirical probability measure restricted to Borel σ-algebra of $\mathcal{W}$ is a Borel probability measure.

Unknown regression coefficients are estimated by an ε_t-M-estimator based on a loss function defined by the formula

$$\mathsf{L}(\mu\,;\beta) = \int \rho(y - x^\top\beta)\mu(\mathrm{d}y, \mathrm{d}x). \tag{5}$$

Especially, for empirical distribution of observations we receive

$$\begin{aligned}
&\mathsf{L}(\mathcal{E}_t(\bullet \,|\, (y_1, x_1), (y_2, x_2), \ldots, (y_t, x_t))\,; \beta) = \\
&= \int \rho(y - x^\top\beta)\ \mathcal{E}_t(\mathrm{d}y, \mathrm{d}x \,|\, (y_1, x_1), (y_2, x_2), \ldots, (y_t, x_t)) \\
&= \frac{1}{t}\sum_{i=1}^{t} \rho(y_i - x_i^\top\beta).
\end{aligned}$$

An ε_t-M-estimator is any $\hat{\beta}_t \in \Theta$ fulfilling for all $\beta \in \Theta$

$$\mathsf{L}(\mathcal{E}_t(\bullet \,|\, (Y_1, X_1), (Y_2, X_2), \ldots, (Y_t, X_t))\,;\hat{\beta}_t) < \tag{6}$$
$$< \mathsf{L}(\mathcal{E}_t(\bullet \,|\, (Y_1, X_1), (Y_2, X_2), \ldots, (Y_t, X_t))\,;\beta) + \varepsilon_t.$$

Now, the studied situation is fully described and we are proceeding to assumptions. We introduce the following list of assumptions:

Assumption R1

$\Theta \subset \mathbb{R}^d$ is a closed subset.

Assumption R2

$\varepsilon_t > 0$ for any $t \in \mathbb{N}$ and $\lim\limits_{n \to +\infty} \varepsilon_t = 0$.

Assumption R3

There is a Borel measure ν defined on $\mathbb{R}^{d+1}$ and $\Omega_1 \subset \Omega$ such that $\text{prob}\,(\Omega_1) = 1$ and for all $\omega \in \Omega_1$

$$\mathcal{E}_t(\bullet \,|\, (X_1(\omega), \mathrm{e}_1(\omega)), (X_2(\omega), \mathrm{e}_2(\omega)), \ldots, (X_t(\omega), \mathrm{e}_t(\omega))) \xrightarrow[n \to +\infty]{w} \nu \ .$$

Assumption R4

For any $\beta \in \Theta$

$$\int \rho(e)\,\nu\,(\mathrm{d}x, \mathrm{d}e) \leq \int \rho(e + x^\top(\beta_0 - \beta))\,\nu\,(\mathrm{d}x, \mathrm{d}e)\,.$$

Assumption R5

Function $\rho : \mathbb{R} \to \mathbb{R}$ is nonnegative and continuous.

Assumption R6

There are a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ which is continuous, nondecreasing and $\Omega_2 \subset \Omega$, $\operatorname{prob}(\Omega_2) = 1$ fulfilling:

1. For all $t \in \mathbb{R}$ $\quad \rho(t) \leq \psi(|t|)$.
2. For all $t > 0$ $\quad \int \psi(|e| + t\|x\|)\nu(\mathrm{d}x, \mathrm{d}e) < +\infty$.
3. For all $t > 0$, $\omega \in \Omega_2$

$$\frac{1}{t}\sum_{i=1}^{t} \psi(|\mathrm{e}_i(\omega)| + t\|X_i(\omega)\|) \xrightarrow[n\to+\infty]{} \int \psi(|e| + t\|x\|)\nu(\mathrm{d}x, \mathrm{d}e).$$

Assumption R7

Denoting

$$\begin{aligned} \mathrm{H}_\rho &= \liminf_{\Delta\to+\infty} \inf\left\{\rho(t) \;\middle|\; |t| > \Delta, t \in \mathbb{R}\right\}, \\ \mathrm{M} &= \inf\left\{\nu\left(\left\{(x,e) \in \mathbb{R}^{d+1} \;\middle|\; x^\top\gamma \neq 0\right\}\right) \;\middle|\; \|\gamma\| = 1, \gamma \in \mathbb{R}^d\right\}, \end{aligned}$$

we require $\mathrm{M} > 0$ and a balance

$$\mathrm{H}_\rho \mathrm{M} > \int \rho(e)\,\nu\,(\mathrm{d}x, \mathrm{d}e)\,.$$

Lemma

For any $\Delta > 0$ *we have*

$$\lim_{\kappa\to+\infty} \inf_{\|\gamma\|=1} \nu\left(\left\{(x,e) \;\middle|\; \kappa\left|x^\top\gamma\right| \geq \Delta + |e|\right\}\right) = \mathsf{M}\,.$$

Theorem

If Assumptions R1- R7 are fulfilled.
Then $\hat{\beta}_t$ *exists for any* $t \in \mathbb{N}$ *and for any* $\beta \in \Theta$

$$\mathsf{L}(\mu_{\beta_0}\,;\beta) = \int \rho(e + x^\top(\beta_0 - \beta))\,\nu\,(\mathrm{d}x, \mathrm{d}e)\,. \tag{7}$$

Further, for all $\omega \in \Omega_0 = \Omega_1 \cap \Omega_2$ *the sequence* $\hat{\beta}_t(\omega)$, $t \in \mathbb{N}$ *is compact and*

$$\emptyset \neq \mathsf{Ls}\left(\hat{\beta}_t(\omega)\,,\ t \in \mathbb{N}\right) \subset \mathrm{argmin}\left\{\mathsf{L}(\mu_{\theta_0}\,;\theta) \mid \theta \in \Theta\right\}. \tag{8}$$

Proof.

This theorem is a particular case of Theorem 1.
We set $\mathcal{X} = \mathcal{Y} = \mathbb{R}^{d+1}$, $\Phi = \mathbb{R}^d$.
Further, we set

$$\mathcal{F} = \left\{ (y, x) \mapsto \psi(|y - x^\top \beta_0| + t\|x\|) \mid t > 0 \right\}.$$

Hence, it can be shown that Assumptions A1-A9 are fulfilled. $\square$

- Chen, X.R.; Wu, Y.H.: Strong consistency of *M*-estimates in linear model. J. Multivariate Analysis **27,1**(1988), 116-130.
- Dodge, Y.; Jurečková, J.: *Adaptive Regression.* Springer-Verlag, New York, 2000.
- Hoffmann-Jørgensen, J.: *Probability with a View Towards to Statistics I, II.* Chapman and Hall, New York, 1994.
- Huber, P.J.: *Robust Statistics*. John Wiley & Sons, New York 1981.
- J.Jurečková: Asymptotic representation of M-estimators of location. Math. Operat. Stat. Sec. Stat. **11,1**(1980), 61-73.

- J.Jurečková: Representation of M-estimators with the second-order asymptotic distribution, Statistics & Decision **3**(1985), 263-276.
- J.Jurečková, P.K.Sen: *Robust Statistical Procedures*. John Wiley & Sons, Inc., New York 1996.
- Kelley, J.L.: *General Topology*. D. van Nostrand Comp., New York, 1955.
- Knight, K.: Limiting distributions for L_1-regression estimators under general conditions. Ann. Statist. **26,2**(1998), 755-770.

- Robinson, S.M.: Analysis of sample-path optimization. Math. Oper. Res. **21,3**(1996), 513-528.
- A.M.Leroy, P.J.Rousseeuw: *Robust Regression and Outlier Detection*. John Wiley & Sons, New York 1987.
- Rockafellar, T.; Wets, R.J.-B.: *Variational Analysis*. Springer-Verlag, Berlin, 1998.
- van der Vaart, A.W.; Wellner, J.A.: *Weak Convergence and Empirical Processes*. Springer, New York, 1996.