17. Extremum Estimators

Hayashi pp. 445-486 and 497-500

Extremum Estimators

Extremum estimators are a class of estimatros that include (linear and nonlinear) least squares, (linear and nonlinear) GMM, and ML as special cases.

An estimator $\hat{\boldsymbol{\theta}}$ is called an **extremum estimator** if there is a scalar objective function $Q_n(\boldsymbol{\theta})$ such that

$$\hat{\boldsymbol{\theta}} \text{ maximizes } Q_n(\boldsymbol{\theta}) \quad \text{subject to} \quad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p,$$

where:

- Θ, the **parameter space**, is the set of possible parameter values

- $\mathbb{R}^p$ is finite dimensional Eculidean space

- $Q_n(\boldsymbol{\theta})$ depends not only on $\boldsymbol{\theta}$ but also on the sample/data $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)$, with $\mathbf{w}_t$ the t-th observation and n the sample size.

The dependence of the objective function on the sample of size n is signalled by the subscript n.

Measurablility of $\hat{\theta}$

Lemma 7.1 (Existance of extremum estimators): We suppose that

a. the parameter space Θ is a compact subset of $\mathbb{R}^p$,

b. $Q_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for any data $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)$ and

c. $Q_n(\boldsymbol{\theta})$ is a measurable function of the data for all $\boldsymbol{\theta}$ of the data that solves the maximization problem.

Two Clases of Extremum Estimators

(a) M-Estimators: An extremum estimator is an **M-Estimators** if the objective function is a sample average:

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} m(\mathbf{w}_t; \boldsymbol{\theta}),$$

where: m is a real-valued function of $(\mathbf{w_t}, \boldsymbol{\theta})$.

The **maximum likelihood (ML)** and the **nonlinear least squares (NLS)** are two examples of an M-estimator.

Two Clases of Extremum Estimators (cont'd)

(b) GMM: An extremum estimator is a **GMM estimator** if the objective function can be written as

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{g}_n(\boldsymbol{\theta})'\widehat{\mathbf{W}}\mathbf{g}_n(\boldsymbol{\theta}) \quad \text{with} \quad \underset{(K\times 1)}{\mathbf{g}_n(\boldsymbol{\theta})} \equiv \frac{1}{n}\sum_{t=1}^{n}\mathbf{g}(\mathbf{w_t};\boldsymbol{\theta}),$$

where $\widehat{\mathbf{W}}$ is a $K \times K$ symmetric and positive definite matrix that defines the distance of $\mathbf{g}_n(\boldsymbol{\theta})$ from zero, and that can depend on the data.

Note: maximizing the GMM objective function above is equivalent to minimizing the distance $\mathbf{g}_n(\boldsymbol{\theta})'\widehat{\mathbf{W}}\mathbf{g}_n(\boldsymbol{\theta})$.

Two Clases of Extremum Estimators (cont'd)

Maximum Likelihood (ML)

A ML estimator is an example of an M-estimator for the case where $\{\mathbf{w}_t\}$ is i.i.d.

The ML model is a set of i.i.d. sequences $\{\mathbf{w}_t\}$ where the density of $\mathbf{w}_t$ is a member of the family of densities indexed by a finite-dimensional vector $\boldsymbol{\theta} : f(\mathbf{w}_t; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta$.

The ML model is **parametric** since the parameter vector $\boldsymbol{\theta}$ is finite dimensional.

At the true parameter value $\boldsymbol{\theta}_0$, the density of the true DGP is $f(\mathbf{w}_t; \boldsymbol{\theta}_0)$.

The model is **correctly specified** if $\boldsymbol{\theta}_0 \in \Theta$.

ML (cont'd)

Since $\{\mathbf{w}_t\}$ is independently distributed, the joint density of the data $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)$ is

$$f(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n; \boldsymbol{\theta}_0) = \prod_{t=1}^{n} f(\mathbf{w}_t; \boldsymbol{\theta}_0).$$

The ML estimator of $\boldsymbol{\theta}_0$ is the $\boldsymbol{\theta}$ that maximizes the likelyhood function

Monotone transformation means that maximizing the likelihood function is equivalent to maximizing the **log likelihood function**:

$$\log f(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n; \boldsymbol{\theta}_0) = \log \left[\prod_{t=1}^{n} f(\mathbf{w}_t; \boldsymbol{\theta}_0) \right] = \sum_{t=1}^{n} \log\, f(\mathbf{w}_t; \boldsymbol{\theta}).$$

ML(cont'd)

The ML estimator of $\boldsymbol{\theta}_0$, therefore, is an M-estimator with

$$m(\mathbf{w}_t;\boldsymbol{\theta}) = \log f(\mathbf{w}_t;\boldsymbol{\theta}), \quad \text{that is,} \quad Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \log\, f(\mathbf{w}_t;\boldsymbol{\theta}).$$

Eg. Estimating the mean of a normal distribution:

Let $(w_1, w_2, \ldots, w_n)$ be a scalar i.i.d. sequence with the distribution of w_t given by $N(\mu, \sigma^2)$, so $\boldsymbol{\theta} = (\mu, \sigma^2)'$ and

$$f(w_t;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(w_t-\mu)^2}{2\sigma^2}\right].$$

ML(cont'd)

The average log likelihood of the data $(w_1, w_2, \ldots, w_n)$ is

$$\frac{1}{n}\sum_{t=1}^{n} \log\, f(w_t; \mu, \sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{n}\sum_{t=1}^{n}\left[\frac{(w_t-\mu)^2}{2\sigma^2}\right].$$

Conditional ML

Conditional ML:

Let $f(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}_0)$ be the conditional density of y_t given $\mathbf{x}_t$, and let $f(\mathbf{x}_t; \boldsymbol{\psi}_0)$ be the marginal density of $\mathbf{x}_t$. Then

$$f(y_t, \mathbf{x}_t, \boldsymbol{\theta}_0, \boldsymbol{\psi}_0) = f(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}_0) f(\mathbf{x}_t; \boldsymbol{\psi}_0)$$

is the density of $\mathbf{w}_t = (y_t, \mathbf{x}_t^{'})^{'}$.

The average log likelihood of the data $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)$ is

$$\frac{1}{n}\sum_{t=1}^{n} \log f(\mathbf{w}_t; \boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{1}{n}\sum_{t=1}^{n} \log f(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}) + \frac{1}{n}\sum_{t=1}^{n} \log f(\mathbf{x}_t; \boldsymbol{\psi}).$$

Conditional ML (cont'd)

The **conditional ML estimator** of $\boldsymbol{\theta}_0$ maximizes the (1st term on the right) average log likelihood, and thus constitutes an M-estimator with

$$m(\mathbf{w}_t;\boldsymbol{\theta}) = \log f(y_t \mid \mathbf{x}_t;\boldsymbol{\theta}), \quad \text{that is,} \quad Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \log f(y_t \mid \mathbf{x}_t;\boldsymbol{\theta}).$$

Eg. Probit: Consider a scalar binary variable, $y_t \in \{0,1\}$, and a vector of regressors, $\mathbf{x}_t$. The conditional probability of y_t is given by

$$f(y_t \mid \mathbf{x}_t;\boldsymbol{\theta}_0) = \Phi(\mathbf{x}_t'\boldsymbol{\theta}_0)^{y_t}[1-\Phi(\mathbf{x}_t'\boldsymbol{\theta}_0)]^{1-y_t}.$$

Conditional ML (cont'd)

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution

The ML estimator of $\boldsymbol{\theta}_0$ is an M-estimator with the m function given by

$$m(\mathbf{w}_t; \boldsymbol{\theta}) = \log f(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}) = y_t \log \Phi(\mathbf{x}_t^{'} \boldsymbol{\theta}) + (1 - y_t) \, log[1 - \Phi(\mathbf{x}_t^{'} \boldsymbol{\theta})].$$

Nonlinear Least Squares (NLS)

As before $\mathbf{w}_t = (y_t, \mathbf{x}_t^{'})^{'}$ represents the observation vector partitioned into two groups, y_t and $\mathbf{x}_t$

The model in NLS is a set of stochastic processes $\{y_t, \mathbf{x}_t\}$ such that $E(y_t \mid \mathbf{x}_t)$ is a member of the parametric family of functions $\varphi(\mathbf{x}_t; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

The functional form $\varphi(\cdot;\cdot)$ is known.

If $\boldsymbol{\theta}_0$ is the true parameter value, then $E(y_t \mid \mathbf{x}_t) = \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0)$ for the true DGP $\{y_t, \mathbf{x}_t\}$.

If we define $\varepsilon_t \equiv y_t - E(y_t \mid \mathbf{x}_t)$, then the correctly specified model can be written as

$$y_t = \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0) + \varepsilon_t,\ E(\varepsilon_t \mid \mathbf{x}_t) = 0,\ \boldsymbol{\theta}_0 \in \Theta.$$

Nonlinear Least Squares (cont'd)

least squares is the most widely used estimation method to estimate $\boldsymbol{\theta}_0$ in NLS.

The NLS estimator, which minimizes the sum of squared residuals, is an M-estimator with

$$m(\mathbf{w}_t;\boldsymbol{\theta}) = -[y_t - \varphi(\mathbf{x}_t;\boldsymbol{\theta})]^2, \quad \text{that is} \quad Q_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n}[y_t - \varphi(\mathbf{x}_t;\boldsymbol{\theta})]^2$$

Note: maximization of $Q_n(\boldsymbol{\theta})$ is the same as $minimization$ of the sum of squared residuals.

Linear and Nonlinear GMM

Given a linear equation, $y_t = \mathbf{z}_t^{'}\boldsymbol{\theta}_0 + \varepsilon_t$, and a vector of instruments, $\mathbf{x}_t$, we applied GMM to arrive at the orthogonality (zero-mean) condition

$$E[(\mathbf{x}_t \cdot (y_t - \mathbf{z}_t^{'}\boldsymbol{\theta}_0))] = \mathbf{0}$$

The correctly specified model is a set of ergodic stationary processes $\mathbf{w}_t = (y_t, \mathbf{z}_t^{'}, \mathbf{x}_t^{'})^{'}$ such that the zero-mean condition hold for $\boldsymbol{\theta}_0$ in Θ.

The linear GMM estimator of $\boldsymbol{\theta}_0$ is a GMM estimator with the $\mathbf{g}$ function in the GMM objective function given by

$$\mathbf{g}(\mathbf{x}_t; \boldsymbol{\theta}) \equiv \mathbf{x}_t \cdot (y_t - \mathbf{z}_t^{'}\boldsymbol{\theta}) = \mathbf{x}_t \cdot y_t - \mathbf{x}_t\mathbf{z}_t^{'}\boldsymbol{\theta}.$$

[Hyashi Ch. 3+4 for more details on Linear GMM]

Nonlinear GMM

GMM can be readily applied to nonlinear equaitons.

Nonlinear GMM estimation occurs when the K GMM moment conditions $\mathbf{g}(\mathbf{w}_t, \theta)$ are nonlinear functions of the p model parameters $\boldsymbol{\theta}$.

The moment conditions $\mathbf{g}(\mathbf{w}_t, \theta)$ may be $K \geq p$ nonlinear functions satisfying

$$E[\mathbf{g}(\mathbf{w}_t, \theta_0)] = \mathbf{0}$$

Alternatively, for a response variable y_t, L explanatory variables $\mathbf{z}_t$, and K instruments $\mathbf{x}_t$, the model may define a nonlinear error term ε_t given by:

$$a(y_t, \mathbf{z}_t; \boldsymbol{\theta}_0) = \varepsilon_t.$$

so that

$$E[\varepsilon_t] = E[a(y_t, \mathbf{z}_t; \boldsymbol{\theta}_0)] = \mathbf{0}$$

Nonlinear GMM (cont'd)

Note: if $\mathbf{z}_t$ is endogeneous, then we cannot use nonlinear least squares to estimate θ.

Given that $\mathbf{x}_t$ is orthogonal to ε_t, we could define

$$\mathbf{g}(\mathbf{w}_t, \theta_0) = \mathbf{x}_t \varepsilon_t = \mathbf{x}_t a(y_t, \mathbf{z}_t; \boldsymbol{\theta}_0)$$

so that

$$E[\mathbf{g}(\mathbf{w}_t, \theta_0)] = E[\mathbf{x}_t \varepsilon_t] = E[\mathbf{x}_t a(y_t, \mathbf{z}_t; \boldsymbol{\theta}_0)] = \mathbf{0}$$

defines the GMM orthogonality conditions.

The estimator obtained by setting $\mathbf{g}(\mathbf{w}_t; \boldsymbol{\theta}) = \mathbf{x}_t \cdot a(y_t, \mathbf{z}_t; \boldsymbol{\theta})$ is called a **generallized nonlinear instrumental variables estimator**.

Nonlinear GMM (cont'd)

It is more general than the usual nonlinear iv estimator since conditinal homoskedasticity is not assumed.

In general, the GMM moment equations produce a system of K nonlinear equations in p unknowns.

Identification of θ_0 requires that

$$E[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)] = \mathbf{0}$$
$$E[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta})] \neq \mathbf{0} \text{ for } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

and that the $K \times p$ matrix

$$\mathbf{G} = E\left[\frac{\partial \mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right]$$

Nonlinear GMM (cont'd)

has full column rank p.

The sample moment condition for an arbitrary $\boldsymbol{\theta}$ is given by

$$\mathbf{g}(\boldsymbol{\theta}) = n^{-1} \sum_{t=1}^{n} \mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta})$$

[Case 1] If $K = p$, then $\boldsymbol{\theta}_0$ is just identified and the GMM objective function beceoms

$$J(\boldsymbol{\theta}) = n \ \mathbf{g}_n(\boldsymbol{\theta})' \mathbf{g}_n(\boldsymbol{\theta}),$$

which does not depend on a weight matrix

The corresponding GMM estimator is then given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{arg}min}\ J(\boldsymbol{\theta})$$

and solves

$$\mathbf{g}_n(\hat{\theta}) = \mathbf{0}$$

[Case 2] If $K > p$, then $\boldsymbol{\theta}_0$ is overidentified.

Let $\hat{\mathbf{W}}$ denote a $K \times K$ symmetric and p.d. weight matrix, possibly dependent on the data, such that $\hat{\mathbf{W}} \underset{p}{\rightarrow} \mathbf{W}$ as $n \rightarrow \infty$ with $\mathbf{W}$ symmetric and p.d.

Nonlinear GMM (cont'd)

The GMM estimator of $\boldsymbol{\theta}_0$, denoted $\hat{\boldsymbol{\theta}}(\hat{\mathbf{W}})$, is defined as

$$\hat{\boldsymbol{\theta}}(\hat{\mathbf{W}}) = \underset{\theta}{\text{argmin}}\ J(\boldsymbol{\theta}, \hat{\mathbf{W}}) = n\ \mathbf{g}_n(\boldsymbol{\theta})' \hat{\mathbf{W}} \mathbf{g}_n(\boldsymbol{\theta})$$

The first order conditions are

$$\frac{\partial J(\hat{\boldsymbol{\theta}}(\hat{\mathbf{W}}), \hat{\mathbf{W}})}{\partial \boldsymbol{\theta}} = 2\ \mathbf{G}_n(\hat{\boldsymbol{\theta}}(\hat{\mathbf{W}}))' \hat{\mathbf{W}} \mathbf{g}_n(\hat{\boldsymbol{\theta}}(\hat{\mathbf{W}})) = \mathbf{0}$$

$$\mathbf{G}_n(\hat{\boldsymbol{\theta}}(\hat{\mathbf{W}})) = \frac{\partial \mathbf{g}_n(\hat{\boldsymbol{\theta}}(\hat{\mathbf{W}}))}{\partial \boldsymbol{\theta}'}$$

Nonlinear GMM (cont'd)

The efficient GMM estimator uses $\hat{\mathbf{W}} = \hat{\mathbf{S}}^{-1}$ such that

$$\hat{\mathbf{S}} \underset{p}{\rightarrow} \mathbf{S} = \text{avar}(\sqrt{n}\mathbf{g}_n(\boldsymbol{\theta}_0)).$$

If $\{\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)\}$ is an ergodic-stationary MDS then

$$\mathbf{S} = E[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)']$$

If $\{\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)\}$ is a serially correlated linear process then

$$\mathbf{S} = LRV = \Gamma_0 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j') = \Psi(1)\Sigma\Psi(1)'$$

$$\Gamma_0 = E[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)\mathbf{g}'(\mathbf{w}_t, \boldsymbol{\theta}_0)],\ \Gamma_j = E[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)\mathbf{g}(\mathbf{w}_{t-j}, \boldsymbol{\theta}_0)']$$

Nonlinear GMM (cont'd)

Eg. Student's-t Distribution:

Consider a random sample $y_1, \ldots, y_T$ form a centered Students's-t distribution with θ_0 degrees of freedom with pdf

$$f(y_t; \theta_0) = \frac{\Gamma[(\theta_0 + 1)/2]}{(\pi\theta_0)^{1/2}\Gamma(\theta_0/2)}[1 + (y_t^2/\theta_0)]^{-(\theta_0+1)/2}$$

where $\Gamma(\cdot) =$ gamma function

we want to estimate the degrees of freedom parameter θ_0 by GMM using the following moment conditions

$$E[y_t^2] = \frac{\theta_0}{\theta_0 - 2}$$

Nonlinear GMM (cont'd)

$$E[y_t^4] = \frac{3\theta_0^2}{(\theta_0 - 2)(\theta_0 - 4)}, \ \theta_0 > 4$$

Let $\mathbf{w}_t = (y_t^2, y_t^4)'$ and define

$$\mathbf{g}(\mathbf{w}_t, \theta) = \begin{pmatrix} y_t^2 - \theta/(\theta - 2) \\ y_t^4 - 3\theta^2/(\theta - 2)(\theta - 4) \end{pmatrix}$$

Then $E[\mathbf{g}(\mathbf{w}_t, \theta_0)] = \mathbf{0}$ is the moment condition used for defining the GMM estimator for θ_0.

Nonlinear GMM (cont'd)

In this example $K = 2$ and $p = 1 \Rightarrow \theta_0$ is overidentified

Since we assume random sampling, $\mathbf{g}(\mathbf{w}_t, \theta_0)$ is an i.i.d. process.

Using the sample moments

$$\mathbf{g}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{g}(\mathbf{w}_t, \theta) \quad = \begin{pmatrix} \frac{1}{n}\sum_{t=1}^{n} y_t^2 - \theta/(\theta-2) \\ \frac{1}{n}\sum_{t=1}^{n} y_t^4 - 3\theta^2/(\theta-2)(\theta-4) \end{pmatrix}$$

Nonlinear GMM (cont'd)

The GMM objective function has the form

$$J(\theta) = n \ \mathbf{g}_n(\theta)' \hat{\mathbf{W}} \mathbf{g}_n(\theta)$$

where $\hat{\mathbf{W}}$ is a 2×2 p.d. and symmetric weight matrix, possibly dependent on the data, such that $\hat{\mathbf{W}} \underset{p}{\to} \mathbf{W}$.

The efficient GMM estimator uses the weight matrix $\hat{\mathbf{S}}^{-1}$ so that

$$\hat{\mathbf{S}} \underset{p}{\to} \mathbf{S} = E[\mathbf{g}(\mathbf{w}_t, \theta_0)\mathbf{g}(\mathbf{w}_t, \theta_0)']$$

For example, one could use

$$\hat{\mathbf{S}} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{g}(\mathbf{w}_t, \hat{\theta})\mathbf{g}(\mathbf{w}_t, \hat{\theta})'$$

with $\hat{\theta} \underset{p}{\to} \theta$.

Note:

1. In estimating the model, the restriction $\theta > 4$ should be imposed, which may be done by reparameterization. Define

$$\theta = h(\gamma) = \exp(\gamma) + 4, \quad -\infty < \gamma < \infty$$

Nonlinear GMM (cont'd)

and then estimate γ freely. Given

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \underset{d}{\to} N(0, V)$$

a consistent and asymptotically normal estimate for θ, by Slutsky's theorem and the delta method, is $\hat{\theta} = h(\hat{\gamma}) = \exp(\hat{\gamma}) + 4$ where

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, h^{'}(\gamma_0)^2 \times V)$$

2. Since the pdf of the data is known, the most efficient estimator is the maximum likelihood estimator.

18. Asymptotics of Extremum Estimators

Hayashi pp. 445-486 and 497-500

Introduction

As before (from Hayashi, Ch. 2) the two main concepts in asymptotics relate to $consistency$ and $asymptotic\ normality$.

Some intuition:

Consistency: the more data we get, the closer we get to knowing the truth (or we eventually know the truth)

Asymptotic normality: as we get more and more data, averages of random variables behave like normally distributed random variables.

Example: Establishing consistency and asymptotic normality of an iid random sample $X_1, \ldots, X_N$ with $E(X_i) = \mu$ and var$(X_i) = \sigma^2$.

Introduction (cont'd)

Recall also the main **probability theory tools**:

The probability theory tools for establishing consistency of estimators are:

- Laws of Large Numbers (LLNs)

– A LLN is a result that states the conditions under which a sample average of a random variables converges to a population expectation.
– LLNs concern conditions under which the sequence of sample mean converges either in probability or almost surely
– There are many LLN results (eg. Chebychev's LLN, Kolmongorov's LLN, Markov's LLN)

Introduction (cont'd)

The probability tools for establishing asymptotic normality are:

- Central Limit Theorems (CLTs)

– CLTs are about the limiting behaviour of the difference between a sample mean and its expected value
– There are many CLTs (eg. Lindeberg-Levy CLT, Lindeberg-Feller CLT, Liapounov's CLT)

Asymptotics of Extremum Estimators

Asymptotics of Extremum Estimators follows the same intuition.

However, we need more statistical and mathematical tools to establish *consistency* and *asymptotic normality* of the estimators

Establishing consistency of extremum estimators requires **uniform convergence in probability** and **uniform law of large numbers** or its multivariate version

Establishing asymptotic normality, on the other hand, requires **mean value theorem**.

Consistency

Uniform convergence in probability: the function $\hat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$ if

$$\sup_{\theta \in \Theta} \left| \hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta}) \right| \underset{p}{\to} 0 \quad \text{as} \quad n \to \infty$$

Theorem: if there is a function $Q_0(\boldsymbol{\theta})$ such that

(i) $Q_0(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$,

(ii) Θ is compact,

(iii) $Q_0(\boldsymbol{\theta})$ is continuous,

Consistency (cont'd)

(iv) $\hat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$,

then $\hat{\boldsymbol{\theta}} \underset{p}{\to} \boldsymbol{\theta}_0$.

Uniform convergence in probability can be extended to sequence of vector random functions by requiring uniform convergence for each element

A sequence of vector random functinos $\{\hat{\mathbf{h}}_n(\boldsymbol{\theta})\}$ converges uniformly in probability to a nonrandom function $\{\mathbf{h}_0(\boldsymbol{\theta})\}$ if each element converges uniformly.

Such element-by-element convergence is equivalent to convergence in the norm:

$$\sup \left\| \hat{\mathbf{h}}_n(\boldsymbol{\theta}) - \mathbf{h}_0(\boldsymbol{\theta}) \right\| \underset{p}{\to} 0 \quad as \quad n \to \infty,$$

Consistency (cont'd)

where $\|\cdot\|$ is the Euclidean norm.

Sometimes, especially in the case of simulation-based estimators, it may not be feasible to find the true maximizer $\hat{Q}_n$.

Suppose, instead, we have a "near-maximizer" in the sense that $\hat{\boldsymbol{\theta}}$ satisfies

$$\hat{Q}_n(\hat{\boldsymbol{\theta}}) \geq \sup_{\theta \in \Theta} \hat{Q}_n(\boldsymbol{\theta}) + o_p(1).$$

This ensures that the previous theorem's conclusion continues to hold.

A standard tool to prove uniform convergence in probability for estimators (including MLE, NLS, and GMM, each of which depends on sample averages) is the **uniform law of large numbers**

Consistency (cont'd)

Let $a(z_i, \theta)$ be a function of i.i.d. observations, z_i and the parameter θ, and for a matrix $A = [a_{jk}]$, let $\|A\| = (\sum_{j,K} a_{jk}^2)^{1/2}$ be the Euclidean norm

Lemma (Uniform law of large numbers):

If the data are $i.i.d.$, Θ is compact, $a(z_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability one, and there is $d(z)$ with $\|a(z, \theta)\| \leq d(z)$ for all $\theta \in \Theta$ and $E[d(z)] < \infty$, then

- $E[a(z, \theta)]$ is continuous, and
- $\sup_{\theta \in \Theta} \left\| n^{-1} \sum_{i=1}^{n} a(z_i, \theta) - E[a(z, \theta)] \right\| \underset{p}{\to} 0.$

Consistency (cont'd)

Examples

1. Maximum Likelihood:

Let $z_1, z_2, \ldots, z_n$ be $i.i.d$ with a pdf $f(z \mid \theta_0)$ for $\theta_0 \in \Theta$, then the ML objective function is given by

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log\, f(z_i \mid \theta)$$

[Cf. with our earlier definition of this: $Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^{n} \log\, f(\mathbf{w}_t; \boldsymbol{\theta})$.]

Consistency (cont'd)

We can apply the ULLN directly. Suppose that $\log f(z_i \mid \theta)$ is continuous at each $\theta \in \Theta$ with probability one, and

$$E[\sup_{\theta \in \Theta} |\log f(z_i \mid \theta)|] < \infty.$$

Then, by the ULLN

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} \log f(z_i \mid \theta) - E[\log f(z_i \mid \theta)] \right\| \underset{p}{\to} 0,$$

and $Q(\theta) = E[\log f(z_i \mid \theta)]$ is continuous.

Consistency (cont'd)

A likelihood model is *identified* if the distribution of z_i at θ_0 is different from the distribution at any other θ

Formally, for any $\theta \neq \theta_0$, let

$$\Delta = \{z : \ f(z \mid \theta) \neq f(z \mid \theta_0)\}.$$

Then we require

$$\Pr_{\theta_0}(z_i \in \Delta) > 0.$$

An implication of this is that the ratio $f(z_i \mid \theta)/f(z_i \mid \theta_0)$, regarded as a random variable, is not degenerate at 1.

Consistency (cont'd)

Then, consider

$$
\begin{aligned}
Q(\theta) - Q(\theta_0) &= E[\log f(z_i \mid \theta)] - E[\log f(z_i \mid \theta_0)] \\
&= E\left[\log \frac{f(z_i \mid \theta)}{f(z_i \mid \theta_0)}\right] \\
&< \log E\left[\frac{f(z_i \mid \theta)}{f(z_i \mid \theta_0)}\right] \\
&= \log \int \frac{f(z \mid \theta)}{f(z \mid \theta_0)} f(z \mid \theta_0) dz \\
&= \log \int f(z \mid \theta) dz \\
&= \log 1 = 0.
\end{aligned}
$$

Consistency (cont'd)

The strict inequality follows from **Jensen's inequality**, which holds strictly when the random variable is nonconstant and positive.

Thus, it follows from this that for any $\theta \neq \theta_0$, $Q(\theta) < Q(\theta_0)$

Consistency (cont'd)

2. GMM:

Let

$$\hat{g}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta),$$

and

$$g_0(\theta) = E[g(z_i, \theta)] \ \ (= 0 \text{ at } \theta = \theta_0).$$

then

$$\hat{Q}_n(\theta) = -\hat{g}_n(\theta)' \hat{W} \hat{g}_n(\theta).$$

Suppose that $g(z_i, \theta)$ satisfies the condition for the ULLN. Then

$$\sup_{\theta \in \Theta} \|\hat{g}_n(\theta) - g_0(\theta)\| \underset{p}{\to} 0,$$

and $g_0(\theta)$ is continuous.

Consistency (cont'd)

Assume that $\hat{W} \underset{p}{\to} W$, where W is positive semidefinite and finite. Define

$$Q_0(\theta) = -g_0(\theta)' \hat{W} \hat{g}_0(\theta).$$

This is continuous function of θ, since g_0 is continuous and W is positive semidefinite.

It can then be shown (see NM (1994), p. 2132) that:

$$\sup_{\theta \in \Theta} \left| \hat{Q}_n(\theta) - Q_0(\theta) \right| \underset{p}{\to} 0.$$

Consistency (cont'd)

To show that $Q_0(\theta)$ is uniquely maximized at θ_0,

We have (by assumption),

$$g_0(\theta_0) = E[g(z_i, \theta_0)] = 0.$$

Suppose we can also show that

$$W g_0(\theta) \neq 0 \quad \forall \theta \neq \theta_0.$$

Let $R^{'} R = W$. Then,

$$R g_0(\theta) \neq 0 \quad \forall \theta \neq \theta_0.$$

Therefore

$$Q_0(\theta) = -(R g_0(\theta))^{'} (R g_0(\theta)) < 0 \quad \forall \theta \neq \theta_0.$$

So $Q_0(\theta)$ is uniquely maximized at $\theta = \theta_0$.

Asymptotic Normality

Theorem: Suppose $\hat{\theta}$ maximizes $\hat{Q}_n(\theta)$ over Θ and $\hat{\theta} \underset{p}{\to} \theta_0$, and the following hold:

(i) $\theta_0 \in \text{int}(\Theta)$.

(ii) $\hat{Q}_n(\theta)$ is twice continuously differentiable in a neighbourhood $\mathcal{N}(\theta_0)$ of θ_0.

(iii) $\sqrt{n}\nabla_\theta \hat{Q}_n(\theta) \underset{d}{\to} N(0, \Sigma)$.

(iv) there is a $H(\theta)$, continuous at θ_0, such that

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left\| \nabla_{\theta\theta} \hat{Q}_n(\theta) - H(\theta) \right\| \underset{p}{\to} 0$$

Asymptotic Normality (cont'd)

(v) $H = H(\theta_0)$ is nonsingular.

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, H^{-1}\Sigma H^{-1}).$$

For GMM, we can have a slightly modified version of the theorem

Theorem: Consider a GMM estimator with moment function $g(z_i, \theta)$, and weighting matrix $\hat{W}$. Suppose $\hat{\theta} \underset{p}{\to} \theta_0$, and $\hat{W} \underset{p}{\to} W$, and assume:

(i) $\theta_0 \in \text{int}(\Theta)$.

(ii) $g(z_i, \theta)$ is continuously differentiable in a neighbourhood $\mathcal{N}(\theta_0)$ of θ_0.

Asymptotic Normality (cont'd)

(iii) $E[g(z_i, \theta_0)] = 0$.

(iv) $E[\|g(z_i, \theta_0)\|^2]$ is finite

(v)

$$E[\sup_{\theta \in \mathcal{N}(\theta_0)} \|\nabla_\theta g(z_i, \theta)\|] < \infty.$$

(vi) $G'WG$ is nonsingular, where $G = E[\nabla_\theta g(z_i, \theta)]$.

Asymptotic Normality (cont'd)

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, V),$$

where,

$$S = E[g(z_i, \theta_0)g(z_i, \theta_0)'],$$

and

$$V = avar(\hat{\theta}) = (G'WG)^{-1}G'WSWG(G'WG)^{-1}.$$

The intuition for the result and the variance formula can be shown:

In the GMM problem, the FOC for a minimum is

$$\left[\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta g(z_i, \hat{\theta})\right]' \hat{W} \left[\frac{1}{n}\sum_{i=1}^{n} g(z_i, \hat{\theta})\right] = 0.$$

Asymptotic Normality (cont'd)

Expanding the function $g(z_i, \hat{\theta})$:

$$g(z_i, \hat{\theta}) = g(z_i, \theta_0) + \nabla_\theta g(z_i, \bar{\theta})(\hat{\theta} - \theta_0).$$

Substituting this into the FOC:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g(z_i, \hat{\theta})\right]' \hat{W} \left[\frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g(z_i, \bar{\theta})\right](\hat{\theta} - \theta_0)\right] = 0.$$

Let

$$\hat{G}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g(z_i, \theta_0).$$

Asymptotic Normality (cont'd)

Then we could rearrange the FOC to get

$$\sqrt{n}(\hat{\theta} - \theta_0) = -(\hat{G}_n(\hat{\theta})'\hat{W}\hat{G}_n(\bar{\theta}))^{-1}\hat{G}_n(\hat{\theta})'\hat{W}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(z_i, \theta_0).$$

Then under the condtions of the theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(z_i, \theta_0) \underset{d}{\to} N(0, S),$$

$$\hat{W} \underset{p}{\to} W,$$

Asymptotic Normality (cont'd)

$$\hat{G}_n(\hat{\theta}) \underset{p}{\to} G,$$
$$\hat{G}_n(\bar{\theta}) \underset{p}{\to} G.$$

So

$$-(\hat{G}_n(\hat{\theta})'\hat{W}\hat{G}_n(\bar{\theta}))^{-1}\hat{G}_n(\hat{\theta})'\hat{W} \underset{p}{\to} (G'WG)^{-1}G'W,$$

and by the Slutsky Theorem,

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, (G'WG)^{-1}G'WSWG(G'WG)^{-1}).$$

Asymptotic Normality (cont'd)

As we saw before, $W = S^{-1}$ (i.e., $\hat{W} \underset{p}{\rightarrow} S^{-1}$.) gives the optimal weighting matrix.

So that

$$V = (G'WG)^{-1}G'WSWG(G'WG)^{-1} = (G'S^{-1}G)^{-1}.$$

Example: MLE

Assume the log likelihood function $\log f(z_i \mid \theta)$ is twice continuously differentiable and that $\hat{\theta} \in int(\Theta)$ to the maximum likelihood problem.

Asymptotic Normality (cont'd)

Then, FOC

$$0 = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(z_i \,|\hat{\theta}).$$

Applying the mean-value version of Taylor's theorem to write each individual term in the sum, we get

$$\nabla_\theta \log f(z_i \,|\hat{\theta}) = \nabla_\theta \log f(z_i \,|\theta_0) + \nabla_{\theta\theta} \log f(z_i \,|\bar{\theta})(\hat{\theta} - \theta_0),$$

where $\bar{\theta}$ is between θ_0 and $\hat{\theta}$.

Thus, we can write

$$0 = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(z_i \,|\theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} \log f(z_i \,|\bar{\theta})\right](\hat{\theta} - \theta_0).$$

Asymptotic Normality (cont'd)

Rearranging gives

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_{\theta\theta} \log f(z_i \mid \bar{\theta})\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_{\theta} \log f(z_i \mid \theta_0).$$

Let

$$H = E[\nabla_{\theta\theta} \log f(z_i \mid \theta_0)],$$

and

$$J = E[(\nabla_{\theta} \log f(z_i \mid \theta_0))(\nabla_{\theta} \log f(z_i \mid \theta_0))'].$$

Asymptotic Normality (cont'd)

Then, as long as

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta\theta}\log f(z_i \mid \theta)$$

converges in probability to

$$E[\nabla_{\theta\theta}\log f(z_i \mid \theta)],$$

uniformly in a neighbourhood of θ_0, and since $\bar{\theta} \underset{p}{\to} \theta_0$ (it is "between" $\hat{\theta}$ and θ_0, we have

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta\theta}\log f(z_i \mid \bar{\theta}) \underset{p}{\to} H.$$

Asymptotic Normality (cont'd)

Also, $\nabla_\theta \log f(z_i \mid \theta_0)$ has mean 0 and variance J, so by the CLT,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \log f(z_i \mid \theta_0) \underset{d}{\to} N(0, J).$$

Thus, by the Slutsky Theorem, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to the product of $-H^{-1}$ and a $N(0, J)$ random variable. That is

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, H^{-1} J H^{-1}).$$

Asymptotic Normality (cont'd)

Information equality means that

$$-E[\nabla_{\theta\theta}\text{log}f(z_i \mid \theta_0)] = E[(\nabla_\theta \text{log}f(z_i \mid \theta_0))(\nabla_\theta \text{log}f(z_i \mid \theta_0))'],$$

or

$$-H = J.$$

Thus, the limit distribution simplifies to $N(0, J^{-1})$, where J is the "Fisher information matrix."

Numerical Optimizations

Numerical optimization concerns the computational aspects of extremum estimators.

In the case where there are no closed-form formulae, some numerical algorithm needs to be employed to locate the maximum

Newton-Raphson:

Is probably the best known method for finding iterative approximations to the roots of a real-valued function.

Consider M-estimators with the objective function

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} m(\mathbf{w}_t; \boldsymbol{\theta}),$$

where: m is a real-valued function of $(\mathbf{w_t}, \boldsymbol{\theta})$.

Numerical Optimizations (cont'd)

Since $Q_n(\boldsymbol{\theta})$ is twice continuously differentiable, there is a second-order Taylor expansion given by

$$Q_n(\boldsymbol{\theta}) \cong Q_n(\hat{\boldsymbol{\theta}}_j) + \mathbf{s}_n(\hat{\boldsymbol{\theta}}_j)'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j)'\mathbf{H}_n(\hat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j)$$

where $\hat{\boldsymbol{\theta}}_j$ is the estimate in the j-th round of the iterative procedure, and $\mathbf{s}_n$ and $\mathbf{H}_n$ are the gradient and the Hessian of the objective function:

$$\mathbf{s}_n(\boldsymbol{\theta}) \equiv \frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \quad \mathbf{H}_n(\boldsymbol{\theta}) \equiv \frac{\partial^2 Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}.$$

Numerical Optimizations (cont'd)

The $(j+1)$-th round estimator $\hat{\boldsymbol{\theta}}_{j+1}$ is the maximizer of the quadratic function on the right-hand side of the second-order Taylor expansion above.

Let $\hat{\boldsymbol{\theta}}_j$ be the vector of parameters on the jth iteration, and let $\hat{\boldsymbol{\theta}}_{j+1}$ be the value on the next iteration.

To motivate how we get from $\hat{\boldsymbol{\theta}}_j$ to $\hat{\boldsymbol{\theta}}_{j+1}$, we could use a mean value expansion (row by row) to write

$$\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}}_{j+1}) = \sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}}_j) + \left[\sum_{i=1}^{N} \mathbf{H}_i(\hat{\boldsymbol{\theta}}_j)\right] (\hat{\boldsymbol{\theta}}_{j+1} - \hat{\boldsymbol{\theta}}_j) + \mathbf{r}(\hat{\boldsymbol{\theta}}_j)$$

where $\mathbf{r}(\hat{\boldsymbol{\theta}}_j)$ represents a vector of remainder terms.

Numerical Optimizations (cont'd)

If $\hat{\boldsymbol{\theta}}_{j+1} = \hat{\boldsymbol{\theta}}$, then the left-hand side of the preceding equation becomes zero.

Setting the left-hand side to zero, ignoring $\mathbf{r}(\hat{\boldsymbol{\theta}}_j)$, and assuming that the Hessian evaluated at $\hat{\boldsymbol{\theta}}_j$ is nonsingular, we can get

$$\hat{\boldsymbol{\theta}}_{j+1} = \hat{\boldsymbol{\theta}}_j - \left[\sum_{i=1}^{N} \mathbf{H}_i(\hat{\boldsymbol{\theta}}_j)\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}}_j)\right]$$

This equation provides an iterative method for finding $\hat{\boldsymbol{\theta}}$.

To begin the iteration process, we must choose a vector of starting values $(\hat{\boldsymbol{\theta}}_0)$.

This iterative procedure is called the **Newton-Raphson algorithm**.

Numerical Optimizations (cont'd)

The algorithm often converges quickly to the global maximum if the objective function is concave

For ML estimators, when the golbal maximum $\hat{\boldsymbol{\theta}}$ is thus obtained, the estimate of Avar($\hat{\boldsymbol{\theta}}$) is obtained as $-\mathbf{H}_n(\hat{\boldsymbol{\theta}})^{-1}$.

Numerical Optimizations (cont'd)

Gauss-Newton

Consider GMM estimators with the objective function

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{g}_n(\boldsymbol{\theta})'\widehat{\mathbf{W}}\mathbf{g}_n(\boldsymbol{\theta}) \quad \text{with} \quad \underset{(K\times 1)}{\mathbf{g}_n(\boldsymbol{\theta})} \equiv \frac{1}{n}\sum_{t=1}^{n}\mathbf{g}(\mathbf{w_t};\boldsymbol{\theta}),$$

where $\widehat{\mathbf{W}}$ is a $K \times K$ symmetric and positive definite matrix that defines the distance of $\mathbf{g}_n(\boldsymbol{\theta})$ from zero, and that can depend on the data.

Numerical Optimizations (cont'd)

The first-order Taylor expansion of $\mathbf{g}_n(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}$ can be given by

$$\begin{aligned}\underset{(K\times 1)}{\mathbf{g}_n(\boldsymbol{\theta})} &\cong \mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) + \underset{(K\times p)}{\mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)}\underset{(p\times 1)}{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j)} \\ &= [\mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) - \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)\boldsymbol{\theta}_j] - [-\mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)]\boldsymbol{\theta} \\ &= \mathbf{v}_j - \mathbf{G}_j\boldsymbol{\theta},\end{aligned}$$

where

$$\mathbf{v}_j \equiv \mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) - \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)\boldsymbol{\theta}_j, \mathbf{G}_j \equiv -\mathbf{G}_n(\hat{\boldsymbol{\theta}}_j).$$

[Note: $\mathbf{G}_n(\boldsymbol{\theta}) \equiv \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$]

Numerical Optimizations (cont'd)

If the $\mathbf{g}_n(\boldsymbol{\theta})$ function in the expression for the GMM objective function were the linear function $\mathbf{v}_j - \mathbf{G}_j\boldsymbol{\theta}$, then the objective function would be quadratic in $\boldsymbol{\theta}$ and the maximizer (or the minimizer of the GMM distance) would be the linear GMM estimator given by

$$\hat{\boldsymbol{\theta}}_{j+1} = (\mathbf{G}_j^{'}\widehat{W}\mathbf{G}_j)^{-1}\mathbf{G}_j^{'}\widehat{W}\mathbf{v}_j,$$

which is the $(j+1)$-th iteration in the **Gauss-Newton algorithm**.

Unlike in Newton-Raphson method, there is no need to calculate second-order derivatives in the case of the Gauss-Newton method.