GMM, Indirect Inference and Bootstrap

Maximum Likelihood

Willi Mutschler

TU Dortmund

Winter 2015/2016

Maximum likelihood
Basic idea

- The basic idea is very natural:
- Choose the parameters such that the probability (likelihood) of the observations $x_1, \ldots, x_n$ as a function of the unknown parameters $\theta_1, \ldots, \theta_r$ is maximized
- **Likelihood function**

$$L(\theta; x_1, \ldots, x_n) = \begin{cases} P(X_1 = x_1, \ldots, X_n = x_n; \theta) \\ f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) \end{cases}$$

Maximum likelihood

Basic idea

- For simple random samples

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

- Maximize the likelihood

$$L(\hat{\theta}; x_1, \ldots, x_n) = \max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n)$$

- ML estimate $\hat{\theta} = \arg\max L(\theta; x_1, \ldots, x_n)$
- ML estimator $\hat{\theta} = \arg\max L(\theta; X_1, \ldots, X_n)$

Maximum likelihood

Basic idea

- Because sums are easier to deal with than products, and because sums are subject to limit laws, it is common to maximize the **log-likelihood**

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f_X(X_i; \theta)$$

- The ML estimator is the same as before, since

$$\begin{aligned} \hat{\theta} &= \arg\max \ln L(\theta; X_1, \ldots, X_n) \\ &= \arg\max L(\theta; X_1, \ldots, X_n) \end{aligned}$$

- Further numerical issues: densities are very small! Solution: it is advisable to multiply it with a factor, ML estimator is unchanged.

- Usually, we find $\hat{\theta}$ by solving the system of equations

$$\begin{aligned}\partial \ln L/\partial\theta_1 &= 0\\ &\vdots\\ \partial \ln L/\partial\theta_r &= 0\end{aligned}$$

- The gradient vector $g(\theta) = \partial \ln L(\theta)/\partial\theta$ is called **score vector** or score
- If the log-likelihood is not differentiable other maximization methods must be used

Maximum likelihood

Example

- Let $X \sim Exp(\lambda)$ with density $f(x;\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x;\lambda) = 0$ else
- Likelihood of i.i.d. random sample

$$L(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

- Log-likelihood

$$\ln L(\lambda; x_1, \ldots, x_n) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

- Set the derivative to zero

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i \stackrel{!}{=} 0,$$

hence

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

- The ML estimator for λ is

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

consistent but biased!

- The log-likelihood and the score vector are

$$\begin{aligned} \ln L(\theta) &= \sum_{i=1}^{n} \ln f_X(X_i;\theta) \\ \frac{\partial \ln L(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \frac{\partial \ln f_X(X_i;\theta)}{\partial \theta} \end{aligned}$$

- The contributions $\ln f_X(X_i;\theta)$ are random variables
- The contributions $\partial \ln f_X(X_i;\theta)/\partial\theta$ are random vectors
- Hence, limit laws can be applied to the (normalized) sums

- For all θ

$$\begin{aligned} \int e^{\ln L(\theta)} dx &= \int L(\theta; x_1, \ldots, x_n)\, dx \\ &= 1 \end{aligned}$$

since $L(\theta)$ is a joint density function of $X_1, \ldots, X_n$

- Define the matrix $G(\theta, X_1, \ldots, X_n)$ of gradient contributions

$$G_{ij}(\theta, X_i) = \frac{\partial \ln f_X(X_i;\theta)}{\partial \theta_j}$$

- The column sums are the gradient vector with elements

$$g_j(\theta) = \sum_{i=1}^{n} G_{ij}(\theta, X_i)$$

- The expected gradient vector is $E_\theta(g(\theta)) = 0$ [P]

- The covariance matrix of gradient vector

$$Cov\left(g\left(\theta\right)\right) = E\left(g\left(\theta\right)g\left(\theta\right)'\right)$$

is called information matrix (and often denoted $\mathcal{I}(\theta)$)

- Information matrix equality [P]

$$\begin{aligned} Cov\left(g\left(\theta\right)\right) &= -E\left(H\left(\theta\right)\right) \\ Cov\left(\frac{\partial \ln L\left(\theta\right)}{\partial\theta}\right) &= -E\left(\frac{\partial^2 \ln L\left(\theta\right)}{\partial\theta\partial\theta'}\right) \end{aligned}$$

1. Equivariance: If $\hat{\theta}$ is the ML estimator for θ, then $h(\hat{\theta})$ is the ML estimator for $h(\theta)$
2. Consistency:
$$plim \hat{\theta}_n = \theta$$
3. Asymptotic normality:
$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} U \sim N\left(0, V\left(\theta\right)\right)$$
4. Asymptotic efficiency: $V\left(\theta\right)$ is the Cramér-Rao bound
5. Computability (analytical or numerical); the covariance matrix of the estimator is a by-product of the numerical method

Equivariance:

- Let $\hat{\theta}$ be the ML estimator of θ
- Let $\psi = h(\theta)$ be a one-to-one function of θ with inverse $h^{-1}(\psi) = \theta$
- Then the ML estimator of ψ satisfies

$$\frac{d \ln L\left(h^{-1}(\psi)\right)}{d\psi} = \frac{d \ln L(\theta)}{d\theta} \frac{dh^{-1}(\psi)}{d\psi} = 0$$

which holds at $\hat{\psi} = h\left(\hat{\theta}\right)$

Consistency

- The parameter θ is **identified** if for all $\theta' \neq \theta$ and data $x_1, \ldots, x_n$

$$\ln L\left(\theta' | x_1, \ldots, x_n\right) \neq \ln L\left(\theta | x_1, \ldots, x_n\right)$$

- The parameter θ is **asymptotically identified** if for all $\theta' \neq \theta_0$

$$plim \frac{1}{n} \ln L\left(\theta'\right) \neq plim \frac{1}{n} \ln L\left(\theta_0\right)$$

where θ_0 is the true value of the parameter [P]

Asymptotic normality

- By definition, the ML estimator satisfies

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

- A first order Taylor series expansion of g around the true parameter vector θ_0 gives [P]

$$g(\hat{\theta}) = g(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0) + rest$$

- The (approximate) covariance matrix of $\hat{\theta}$ is

$$Cov(\hat{\theta}) = -\left[E\left(H\left(\theta_0\right)\right)\right]^{-1} = -\left[E\left(\frac{\partial^2 \ln L(\theta_0)}{\partial\theta_0 \partial\theta_0'}\right)\right]^{-1}$$

- A consistent estimator of $Cov(\hat{\theta})$ is

$$\widehat{Cov}(\hat{\theta}) = -\left[H(\hat{\theta})\right]^{-1} = -\left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'}\right)^{-1}$$

- Often, $H(\hat{\theta})$ is a by-product of numerical optimization

- An alternative consistent covariance matrix estimator is

$$\widehat{Cov}(\hat{\theta}) = \left[G(\hat{\theta}; X_1, \ldots, X_n)' G(\hat{\theta}; X_1, \ldots, X_n) \right]^{-1}$$

- This estimator is called outer-product-of-the-gradient (OPG) estimator
- Advantage: Only the first derivatives are required
- Disadvantage: Less reliable in small samples

- Numerical estimation of the parameters of $N(\mu, \sigma^2)$
- Let $X_1, \dots, X_{50}$ be a random sample from $X \sim N(\mu, \sigma^2)$ with $\mu = 5$ and $\sigma^2 = 9$
- Density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)$$

- Log-likelihood function $\ln L\left(\mu, \sigma^2\right) = \sum_{i=1}^{n} \ln f_X(x_i)$

- See `numnormal.R`
- Point estimates

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} 3.64025 \\ 6.90869 \end{pmatrix}$$

- Estimated covariance matrix derived numerically from $H(\hat{\theta})$

$$\widehat{Cov}\left(\hat{\mu}, \hat{\sigma}^2\right) = \begin{pmatrix} 0.13817 & -0.00016 \\ -0.00016 & 1.90918 \end{pmatrix}$$

Maximum likelihood

Example

- See `numnormal.R`
- Point estimates

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} 3.64025 \\ 6.90869 \end{pmatrix}$$

- Estimated covariance matrix derived from theory

$$\widehat{Cov}\left(\hat{\mu}, \hat{\sigma}^2\right) = \begin{pmatrix} 0.13817 & 0 \\ 0 & 1.90920 \end{pmatrix}$$

Maximum likelihood

Example of violated regularity conditions

- Let X be uniformly distributed on the interval $[0, \theta]$
- The density function is

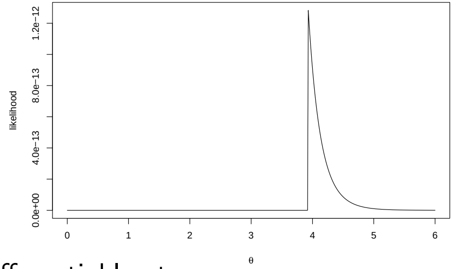
$$f_X(x) = \begin{cases} 1/\theta & \text{for } 0 \leq x \leq \theta \\ 0 & \text{else} \end{cases}$$

- The likelihood function is

$$L(\theta|x_1, \ldots, x_n) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{for } \theta \geq \max_i x_i \\ 0 & \text{else} \end{cases}$$

Maximum likelihood

Example of violated regularity conditions



- $L(\theta)$ is not differentiable at $\max_i x_i$
- Maximum is at $\hat{\theta} = \max_i x_i$
- The estimator is consistent but not asymptotically normal
- Illustration in R

- Maximum likelihood estimation is still possible if the observations are dependent
- The joint density of the observations

$$f_{X_1,\ldots,X_T}(x_1,\ldots,x_T)$$

can be factorized as

$$f_{X_1}(x_1)\cdot\prod_{t=2}^{T} f_{X_t|X_1=x_1,\ldots,X_{t-1}=x_{t-1}}(x_t)$$

- Loglikelihood

$$\ln L = \ln f_{X_1}(x_1) + \sum_{t=2}^{T} \ln f_{X_t|X_1=x_1,\ldots,X_{t-1}=x_{t-1}}(x_t)$$

- If T is large, one may ignore $\ln f_{X_1}(x_1)$
- Computing the loglikelihood is straightforward if

$$f_{X_t|X_1=x_1,\ldots,X_{t-1}=x_{t-1}}(x_t) = f_{X_t|X_{t-1}=x_{t-1}}(x_t)$$