476 Statistics, Spring 2022.

Lecturer: Sergey Nadtochiy.

Lecture 5. Moment-matching estimators and maximum likelihood estimators (MLEs). (Sections 9.6–9.8)

In practice, whenever the statistical model or the unknown parameter θ is complicated (e.g., if the observations are not independent but have inter-dependence, or if each observation comes from a non-standard distribution, or if each observation is a large random vector whose covariance matrix depends on θ, etc.), we have little chance to express the likelihood function L in a sufficiently explicit form to be able to find a Fisher-Neyman decomposition and MVUE. However, we may still be able to approximate numerically (by sampling from the distribution, or by solving a differential equation, etc.) the values of the joint density $L(\cdot;\theta)$ or of the moments $m_k(\theta) := \mathbb{E}^\theta Y_i^k$ for each possible $\theta \in \Theta$. It turns out that this suffices to construct good (although not always best) estimators.

1 Moment-matching estimators (aka method of moments)

This is the oldest method in statistics. Consider a sample of i.i.d. r.v.'s $Y_1, \dots, Y_n$, with moments

$$m_k(\theta) := \mathbb{E}^\theta Y_i^k, \quad k = 1, 2, \dots$$

Def 1. *A moment-matching estimator of θ, of order j, is a statistic $\hat{\theta}$ that satisfies*

$$m_k(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n Y_i^k, \quad k = 1, 2, \dots, j. \tag{1}$$

Ex 1. *Sample mean is a moment-matching estimator of the true mean, of order $j = 1$.*

In practice, we first compute the function $m_k(\cdot)$ (explicitly or via numerical approximation). Then, we compute the right hand side of (1) for the realized sample values $(y_1, \dots, y_n)$ and find (explicitly or via numerical approximation) the value of $\hat{\theta}$ that (approximately) satisfies (1).

Ex 2. *(Optional.) Assume that $Y_i = \max(X_i^1, \dots, X_i^{100})$, where $\{X_i^s\}$ are independent $Exp(1/\theta)$. We do not have an explicit formula for the distribution of Y_i, nor for its moments. However, we can approximate $m_k(\theta) := \mathbb{E}^\theta Y_1^k$ numerically, e.g., by generating 10^4 realizations $(x_m^1(\theta), \dots, x_m^{100}(\theta))$, for $m = 1, 2, \dots, 10^4$ of vectors of 100 independent $Exp(1/\theta)$ random variables, for $\theta = 0.1, 0.2, \dots, 9.9, 10$. Then, we can approximate*

$$m_k(\theta) \approx \frac{1}{10^4}\sum_{m=1}^{10^4} \max(x_m^1(\theta), \dots, x_m^{100}(\theta)), \quad \theta = 0.1, 0.2, \dots, 9.9, 10,$$

and interpolate these values to obtain an approximation of $m_k(\theta)$ for all $\theta \in (0, 10)$ and $k = 1, \dots, j$. Then, we consider the actually observed values of our sample $y_1, \dots, y_n$ and, e.g. find $\hat{\theta}$ that minimizes

$$\sum_{k=1}^j \left[m_k(\hat{\theta}) - \frac{1}{n}\sum_{i=1}^n y_i^k \right]^2.$$

In the exercises we solve in this course, for simplicity, we assume that $m_k(\cdot)$ can be computed explicitly and that the equations (1) can be solved explicitly too. The above discussion of (practically relevant) implementation details is meant just for your information.

Thm 1. *If the mapping $\theta \mapsto (m_1(\theta), \ldots, m_j(\theta))$ has a continuous inverse, then a moment-matching estimator $\hat{\theta}$ of θ, of order j, is* **consistent**.

Proof:

Denote the true value of θ by θ_0 and assume for simplicity that $j = 1$. Then, using (1), LLN, and the continuity of $m_1^{-1}(\cdot)$, we obtain

$$\lim_{n\to\infty} \hat{\theta}_n = \lim_{n\to\infty} m_1^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = m_1^{-1}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = m_1^{-1}\left(m_1(\theta_0)\right) = \theta_0.$$

■

However, a moment-matching estimator may not be the best one, in terms of its MSE.

Ex 3. *Consider a sample of i.i.d. r.v.'s $Y_1, \ldots, Y_n$ from $Gamma(\alpha, \beta)$ distribution, with the pdf*

$$f(y) = e^{-y/\beta}\frac{y^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}, \quad y \geq 0.$$

Recall the first two moments of this distribution:

$$m_1(\alpha, \beta) = \alpha\beta, \quad m_2(\alpha, \beta) = \alpha\beta^2 + \alpha^2\beta^2.$$

The (consistent) moment-matching estimators of (α, β) are given by solving

$$\hat{\alpha}\hat{\beta} = \bar{Y}, \quad \hat{\alpha}\hat{\beta}^2 + \hat{\alpha}^2\hat{\beta}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2,$$

which gives

$$\hat{\alpha} = \frac{n\bar{Y}^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \quad \hat{\beta} = \frac{\bar{Y}}{\hat{\alpha}}.$$

On the other hand,

$$L(y_1, \ldots, y_n; \alpha, \beta) = e^{-\frac{1}{\beta}\sum_{i=1}^{n} y_i}\left(\prod_{i=1}^{n} y_i\right)^{\alpha-1}\beta^{-n\alpha}\Gamma(\alpha)^{-n},$$

which yields (via Fisher-Neyman decomposition) that $(\bar{Y}, \prod_{i=1}^{n} Y_i)$ is a complete sufficient statistic for (α, β). Notice that $(\hat{\alpha}, \hat{\beta})$ are not functions of $(\bar{Y}, \prod_{i=1}^{n} Y_i)$, hence these estimators cannot be the most efficient ones.

2 Maximum likelihood estimator (MLE)

Assume that we can compute the likelihood function numerically but cannot write it down explicitly to find a Fisher-Neyman decomposition. Then we may still be able to compute (approximately) the value of the **maximum likelihood estimator (MLE)**.

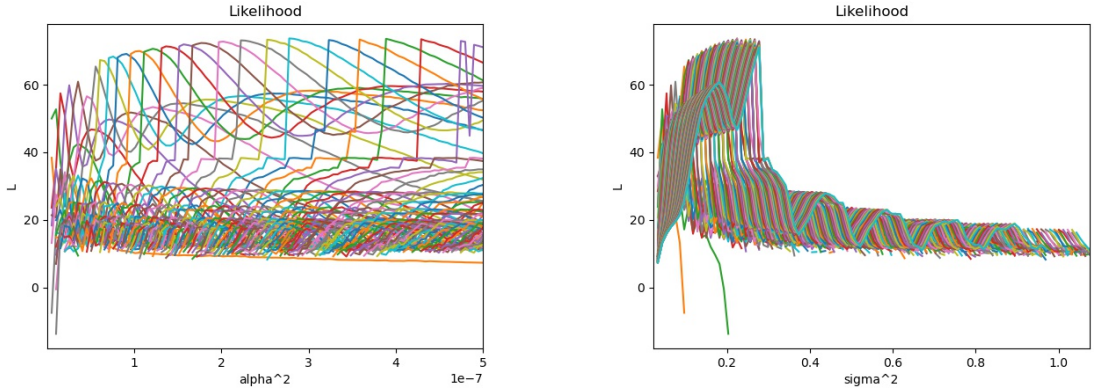

Figure 1: Cross-sectional plots of $\frac{1}{n}\log L(\cdot)$ for INTC ticker computed on Nov 3, 2014. Left: $\frac{1}{n}\log L(\cdot)$ as a function of α^2 (for each fixed value of σ^2). Right: $\frac{1}{n}\log L(\cdot)$ as a function of σ^2 (for each fixed value of α^2).

Def 2. *A statistic $\hat{\theta}$ is a MLE if it attains the maximum*

$$\max_{\theta\in\Theta} L(Y_1,\ldots,Y_n;\theta).$$

Ex 4. *(Optional - for fun. Based on the research article "Consistency of MLE for partially observed diffusions" by S. Nadtochiy and Y. Yin) Assume that the observations $Y_1,\ldots,Y_n$ are not independent and, instead, Y_{i+1} is obtained from Y_i via*

$$Y_{i+1} = G(Y_i, X_{i+1};\theta),$$

where G is a known function (which nevertheless depends on an unknown parameter θ) and $\{X_i\}$ are independent random variables (representing random noise). Then, the system (Y_i) is called a Markov process (or Markov chain).

For a general Markov process, it is unlikely that one can compute explicitly the joint density

$$L(y_1,\ldots,y_n;\theta),$$

which is the likelihood function for a Markov process. It is even less likely that L is available explicitly for a partially observed Markov process (aka hidden Markov model), where every Y_i is a two-dimensional vector, and only one of its components is observed.

Nevertheless, even for partially observed Markov models (relevant in many applications) there exist numerical algorithms for approximating L. Figure 1 shows the normalized log-likelihood for a hidden Markov model used in Finance. This model has two unknown parameters: α^2 and σ^2.

The first important property of MLE is its **consistency**.

Thm 2. *Let $Y_1,\ldots,Y_n$ be i.i.d. random variables and let f denote the pdf or probability function of Y_i. Assume:*

1. *$f(\cdot;\theta) \neq f(\cdot;\theta')$ for $\theta \neq \theta'$.*
2. *$f(y;\theta)$ is continuous in θ for any y.*

3. *The range of possible θ, denoted Θ, is compact.*

4. *There exists $g(y)$, s.t. $\mathbb{E}^{\theta_0} g(Y) < \infty$ for the true θ_0, and*

$$|\log f(y;\theta)| \le g(y), \quad \theta \in \Theta.$$

Then, MLE of θ is consistent.

Proof:

Assume that f is a pdf and denote by θ_0 the true value of θ. Then, LLN yields:

$$\frac{1}{n}\log L(Y_1,\dots,Y_n;\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(Y_i;\theta) \to \mathbb{E}^{\theta_0}\log f(Y_i;\theta).$$

as $n \to \infty$.

Under the assumptions of the theorem one can show that the above convergence also implies that any maximizer $\hat{\theta}_n$ of $L(Y_1,\dots,Y_n;\cdot)$ (and hence of $\frac{1}{n}\log L(Y_1,\dots,Y_n;\cdot)$) converges to a maximizer of $\mathbb{E}^{\theta_0}\log f(Y_i;\cdot)$.

Thus, it suffices to show that θ_0 is the only maximizer of $\mathbb{E}^{\theta_0}\log f(Y_i;\cdot)$. To this end, we notice that

$$\mathbb{E}^{\theta_0}\log f(Y_i;\theta) - \mathbb{E}^{\theta_0}\log f(Y_i;\theta_0) = \mathbb{E}^{\theta_0}\log\frac{f(Y_i;\theta)}{f(Y_i;\theta_0)}$$

$$\le \mathbb{E}^{\theta_0}\left(\frac{f(Y;\theta)}{f(Y;\theta_0)} - 1\right) = \int\left(\frac{f(y;\theta)}{f(y;\theta_0)} - 1\right) f(y;\theta_0)dy = 0,$$

where we used $\log x \le x - 1$ for $x > 0$. It only remains to notice that '$\ge$' in the above can only become '$=$' if $f(\cdot;\theta) = f(\cdot;\theta_0)$, which is only possible if $\theta = \theta_0$, by the assumption of the theorem. ∎

In this course, you will **not** need to verify the assumptions of the above theorem.

Ex 5. *Let $Y_1,\dots,Y_n$ be i.i.d. random variables with distribution $N(\mu,\sigma^2)$. Assume that σ^2 is known while $\mu \in \mathbb{R}$ is not.*

Q 1. *Find a MLE for μ.*

$$L = (2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

To find μ that maximizes the above, we take log *and set its derivative to zero:*

$$0 = \frac{d}{d\mu}\log L = -\frac{1}{2\sigma^2}\frac{d}{d\mu}\sum_{i=1}^{n}(y_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}y_i - \frac{n}{\sigma^2}\mu.$$

Thus, a MLE is given by

$$\hat{\mu} = \bar{Y}.$$

Ex 6. *Let $Y_1,\dots,Y_n$ be i.i.d. random variables with distribution $N(\mu,\sigma^2)$. Assume that both μ and σ^2 are unknown.*

Q 2. *Find a MLE for (μ,σ^2).*

$$L = (2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

To find (μ, σ^2) that maximize the above, we take log *and set its derivatives in μ and σ^2 to zero:*

$$0 = \frac{d}{d\mu}\log L = -\frac{1}{2\sigma^2}\frac{d}{d\mu}\sum_{i=1}^{n}(y_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}y_i - \frac{n}{\sigma^2}\mu,$$

$$\mu = \bar{Y},$$

$$0 = \frac{d}{d\sigma^2}\log L = -\frac{n}{2}\frac{d}{d\sigma^2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\frac{d}{d\sigma^2}\frac{1}{\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\frac{1}{\sigma^4},$$

$$\sigma^2 = \frac{1}{n}(y_i-\bar{y})^2.$$

Note that the MLE we constructed for σ^2 is not exactly equal to (the most efficient) S^2, but the difference between the two is of the order C/n.

Ex 7. *Let $Y_1, \dots, Y_n$ be i.i.d. random variables with distribution $U(0, \theta)$. Assume that θ is unknown.*

Q 3. *Find a MLE for $\theta > 0$.*

$$f(y;\theta) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(y) = \begin{cases} 1/\theta, & y \in [0,\theta], \\ 0, & \textit{else}, \end{cases}$$

$$L = \theta^{-n}\mathbf{1}_{[0,\theta]}(y_1)\cdots\mathbf{1}_{[0,\theta]}(y_n) = \begin{cases} \theta^{-n}, & y_1, \dots, y_n \in [0,\theta], \\ 0, & \textit{else}. \end{cases}$$

Notice that $L = 0$ for any $\theta < y_{(n)}$. On the other hand, L is decreasing for $\theta \geq y_{(n)}$. Thus, a MLE is given by

$$\hat{\theta} = Y_{(n)}.$$

Recall that $Y_{(n)}$ is a complete sufficient statistic, and its bias is of the order C/n.

Exercise 1. *Show that $\bar{Y}$ is a MLE for the unknown success probability p of a Bernoulli sample.*

Ex 8. *(Optional – for fun. Based on a research project by W. Roller.) In this example, we estimate the level of interconnectivity in a financial market. The observations $Y_1, \dots, Y_n$ consist of (normalized) daily returns of $M = 786$ publicly traded stocks: i.e., $Y_i = (Y_i^1, \dots, Y_i^M)$ and Y_i^j is the return of j-th stock on day i.*

We assume that every Y_i is a random **normal vector***, with zero mean and with the covariance matrix*

$$\Sigma_{jk} = \textit{Cov}(Y_i^j, Y_i^k) = \begin{cases} 1, & j = k, \\ \alpha, & j \neq k, \end{cases}$$

where α is the unknown correlation between the returns of different stocks.

Q 4. *Find a MLE for α.*

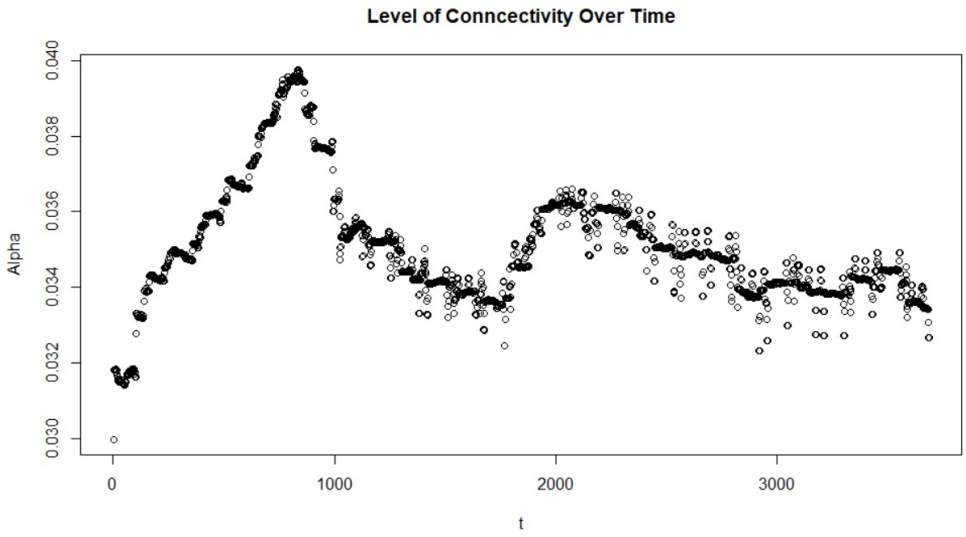

Figure 2: MLE for α.

By analyzing the likelihood function L, it can be shown that MLE $\hat{\alpha}$ is one of the roots of the polynomial

$$a\alpha^3 + b\alpha^2 + c\alpha + d,$$

where

$$a = -\frac{1}{2}nM(M-1)^2,$$

$$b = \frac{1}{2}nM(M-2)(M-1) - \frac{1}{2}(M-2)(M-1)\sum_{i=1}^{n}\sum_{j=1}^{M}(Y_i^j)^2 + \frac{1}{2}(M-1)\sum_{i=1}^{n}\sum_{j,k=1,j\neq k}^{M} Y_i^j Y_i^k,$$

$$c = \frac{1}{2}nM(M-1) - (M-1)\sum_{i=1}^{n}\sum_{j=1}^{M}(Y_i^j)^2,$$

$$d = \frac{1}{2}\sum_{i=1}^{n}\sum_{j,k=1,j\neq k}^{M} Y_i^j Y_i^k.$$

Using the past $n = 250$ days as a sample, we compute $\hat{\alpha}$ for each day in around 15 years (approximately $2004 - -2019$) and plot it on Figure 2.

The fact that MLE is typically a function of a complete sufficient statistic is not surprising – it follows from Fisher-Neyman decomposition:

$$L(Y_1, \ldots, Y_n; \theta) = g(U(Y_1, \ldots, Y_n); \theta)h(Y_1, \ldots, Y_n).$$

Indeed, finding a θ that maximizes L is the same as finding a θ that maximizes $g(U; \theta)$, which depends on the sample only through U, and the latter us typically complete and sufficient. In particular, if MLE is unbiased, it is a MVUE.

MLE does typically have a bias, but the above examples show that the bias of MLE vanishes fast as $n \to \infty$. This is not a coincidence either. In fact, using the Cramer-Rao inequality (which we do not cover in this course and which requires additional technical assumptions on the sample distribution and on the set of parameter values), one can show that

$$\overline{\lim}_{n\to\infty} \text{Eff}(\text{MLE}_n, \hat{\theta}_n) = \overline{\lim}_{n\to\infty} \frac{\text{MSE}(\hat{\theta}_n)}{\text{MSE}(\text{MLE}_n)} \geq 1,$$

for any sequence of unbiased estimators $\hat{\theta}_n$ of θ. Thus, **MLE beats any unbiased estimator asymptotically in terms of its mean squared error**.

Another important property of MLE is its **stability under monotone functional transformations** – aka invariance property.

Thm 3. *Let G be a strictly monotone (i.e., invertible) function and let $\hat{\theta}$ be a MLE for θ. Then, $G(\hat{\theta})$ is a MLE for $G(\theta)$.*

Ex 9. *We know that $\bar{Y}$ is a MLE of the success probability $p \in (0, 1/2]$ of a Bernoulli sample $Y_1, \dots, Y_n$. Then, $\bar{Y}(1-\bar{Y})$ is a MLE for $V(Y_i) = p(1-p)$.*

We can also **use MLE to construct confidence intervals** of any specified asymptotic confidence level. This is done via the following theorem, which states that **MLE is asymptotically normal**.

Thm 4. *Consider i.i.d. random variables $Y_1, \dots, Y_n$, with the pdf (or probability function) of Y_i denoted by $f(\cdot;\theta)$. Let $\hat{\theta}$ be a MLE for θ. Then, for any strictly monotone and differentiable function G, under additional technical assumptions, we have:*

$$Z := \frac{\sqrt{-\mathbb{E}^\theta \, \partial^2_{\theta\theta} \log f(Y_i;\theta)}}{|G'(\theta)|} \sqrt{n} \, (G(\hat{\theta}) - G(\theta)) \to N(0,1),$$

where the convergence holds in the sense of distribution.

Rem 1. $-n\mathbb{E}^\theta \partial^2_{\theta\theta} \log f(Y_i;\theta)$ *is called* **Fisher information**. *The asymptotic variance of the error of MLE is equal to the reciprocal of Fisher information. The Cramer-Rao inequality states the variance of any unbiased estimator cannot be less than the reciprocal of Fisher information.*

Ex 10. *Let the sample $Y_1, \dots, Y_n$ consists of i.i.d. Bernoulli random variables with the unknown success probability $p \in (0, 1/2]$.*

Q 5. *Find a confidence interval for $V(Y_i) = p(1-p)$ of asymptotic confidence level $1-\alpha$.*

We start the construction of a pivot with

$$Z = \frac{\sqrt{-\mathbb{E}^p \, \partial^2_{pp} \log f(Y_1, \dots, Y_n; p)}}{|(p(1-p))'|} \sqrt{n} \, (\bar{Y}(1-\bar{Y}) - p(1-p)) \to N(0,1).$$

Note that

$$\partial^2_{pp} \log f(Y_i;p) = \partial^2_{pp}(Y_i \log p + (1-Y_i)\log(1-p)) = -\frac{Y_i}{p^2} - \frac{1-Y_i}{(1-p)^2},$$

$$\mathbb{E}^p \partial^2_{pp} \log f(Y_i;p) = -\frac{1}{p} - \frac{1}{1-p} = -\frac{1}{p(1-p)},$$

$$|(p(1-p))'| = |1-2p|,$$

$$Z = \frac{1}{\sqrt{p(1-p)(1-2p)^2}}\sqrt{n}\,(\bar{Y}(1-\bar{Y}) - p(1-p)) \to N(0,1).$$

Using Slutsky's theorem, we obtain the pivot

$$U := \frac{1}{\sqrt{\bar{Y}(1-\bar{Y})(1-2\bar{Y})^2}}\sqrt{n}\,(\bar{Y}(1-\bar{Y}) - p(1-p)) \to N(0,1).$$

Performing the standard algorithm we obtain the confidence interval

$$\left[\bar{Y}(1-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})(1-2\bar{Y})^2}{n}},\ \bar{Y}(1-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})(1-2\bar{Y})^2}{n}}\right]$$

for $p(1-p)$, which is of asymptotic confidence level $1-\alpha$.