EDITORIAL

This special issue is devoted to the application of the Maximum Entropy Principle (MaxEnt), especially in econometrics. In a sense, MaxEnt is a nonparametric method for estimation of probability density functions, consistent with data and prior information. It is one principle among other estimation principles in statistics.

Entropy is a quantitative measure of the uncertainty of a stochastic system.The terminology "entropy" in Shannon's work was inspired from the famous formula given by L. Boltzmann (Vorlesungen uber Gastheorie, Leipzig, 1895-1898)

$$H = -\int\int\int f(u,v,w)[\log f(u,v,w)]dudvdw$$

to define the "entropy" of a gas, when the velocity of the molecules is distributed with probability density f. (in Statistical Mechanics).

The MaxEnt principle was rationalized by E. T. Jaynes (1982) as an estimation principle. It was introduced into econometrics by the works of A. Golan, G. Jude and D. Miller (1996) and G. Judge and R. Mittelhammer (2012).

What is a "principle?". Roughly speaking, a principle is a guide to a plausible way of doing things. It is not a *law*, such as the law of large numbers, or the power law, which asserts that, in most of situations, some fact is true. The best way to explain the notion of principle is to look at some familiar principles in statistical estimation.

a) Maximum Likelihood Principle

Considering the random experiment of tossing a biased coin X with unknown $\theta = P(X = 1)$. In addition, suppose we know that $\theta \in \Theta = \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\}$. We toss that coin 4 times, look at the outcomes then make an "educated" guess as what is the most likely value of θ among $\{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\}$. The density of X is

$$f(x,\theta) = \theta^x(1-\theta)^{1-x}1_{\{0,1\}}(x)$$

For a given $\theta \in \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\}$, the joint density of the random sample X_1, X_2, X_3, X_4 is

$$g(x_1,x_2,x_3,x_4|\theta) = \prod_{j=1}^{4} f(x_j,\theta) = \theta^{x_1+x_2+x_3+x_4}(1-\theta)^{4-(x_1+x_2+x_3+x_4)}$$

Let's tabulate this probability for each θ:

θ	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	$\sum_{j=1}^{4} x_j$
	$\frac{81}{256}$	$\frac{16}{256}$	$\frac{1}{256}$	0
	$\frac{27}{256}$	$\frac{16}{256}$	$\frac{3}{256}$	1
$g(.\|\theta)$	$\frac{9}{256}$	$\frac{16}{256}$	$\frac{9}{256}$	2
	$\frac{3}{256}$	$\frac{16}{256}$	$\frac{27}{256}$	3
	$\frac{1}{256}$	$\frac{16}{256}$	$\frac{81}{256}$	4

From this table, we see that the probability of observing $\sum_{j=1}^{4} x_j = 0$ is greatest when $\theta = \frac{1}{4}$, hence, in a kind of reverse logic, when $\sum_{j=1}^{4} x_j = 0$, $\theta = \frac{1}{4}$ is a "most likely" value. Thus, we should "guess" (estimate) θ to be $\frac{1}{4}$ when we observe $\sum_{j=1}^{4} x_j = 0$.

By similar reasoning, we estimate θ by $\frac{1}{4}$, when $\sum_{j=1}^4 x_j = 1$. In summary, $\hat{\theta} = \frac{1}{4}$, when $\sum_{j=1}^4 x_j = 0$ or $\sum_{j=1}^4 x_j = 1$; $\hat{\theta} = \frac{2}{4}$ when $\sum_{j=1}^4 x_j = 2$; and $\hat{\theta} = \frac{3}{4}$ when $\sum_{j=1}^4 x_j = 3$ or $\sum_{j=1}^4 x_j = 4$. In other words, $\hat{\theta}$ is the value θ^* in $\Theta = \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\}$ such that

$$g(x_1, x_2, x_3, x_4|\theta^*) = \max_{\theta\in\Theta} g(x_1, x_2, x_3, x_4|\theta)$$

The estimator $\hat{\theta}$ so obtained is called the *maximum likelihood estimator* (MLE) of θ.

This "principle" is applied to general cases (when the parametric form of a density is known) as follows.

Let X have density $f(x, \theta)$ for $x \in \mathbb{R}^m$, $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $X_1, X_2, ..., X_n$ be a random sample from X. The joint density of the X_j $'s$, as a function of θ, is called the *likelihood function*, denoted as

$$L_n(\theta|X_1, X_2, ..., X_n) = \prod_{j=1}^n f(X_j, \theta)$$

The MLE of θ is defined to be a value $\hat{\theta}_n(X_1, X_2, ..., X_n)$, if it exists, for which

$$L_n(\hat{\theta}_n|X_1, X_2, ..., X_n) = \max_{\theta\in\Theta} L_n(\theta|X_1, X_2, ..., X_n)$$

Important remarks

(i) The above concept of estimators came from the plausible observation in our simple example. It does not depend on the existence of moments. It only depends on a specific form of the density function, with finitely dimensional unknown parameters. However, a priori, there is no reasons why this principle should lead to "good" estimators. In other words, with this principle as a guidance to search for estimators (it provides a systematic method for constructing estimators), we still need to find out whether or not MLEs are good (e.g., satisfying desirable properties, such as consistency, asymptotically normal). As we will see, all this depends on more analytic information about the model densities.

(ii) The MLEs are solutions of maximization problems. But maximization problems might not have solutions! i.e., MLEs might not exists! In such cases, we have to look for other methods of estimation. The point is this. In nice situations, MLE is a popular way to find good estimators. But it is by no means that it is a universal method of estimation.

Rather than giving examples of MLEs, we provide a negative one, namely a real situation where we cannot use MLE since simply its MLE does not exist!

A change-point model

Let X have density

$$f(x, \theta) = ae^{-ax}1_{(0\le x\le\tau)}(x) + be^{-a\tau-b(x-\tau)}1_{(x>\tau)}(x)$$

where $x \in \mathbb{R}$, $\theta = (a, b, \tau) \in \Theta = (0, \infty)^3$.

Given a random sample $X_1, X_2, ..., X_n$ from X, let the random variable $W(\tau)$ be the number of $X_j \leq \tau$. The log-likelihood is

$$\log L_n = \sum_{j=1}^{n} 1_{(0 \leq X_j \leq \tau)} \log a - a \sum_{j=1}^{n} X_j 1_{(0 \leq X_j \leq \tau)} + \sum_{j=1}^{n} [1 - 1_{(0 \leq X_j \leq \tau)}][\log b - (a-b)\tau]$$
$$- b \sum_{j=1}^{n} X_j [1 - 1_{(0 \leq X_j \leq \tau)}]$$

can be put in terms of $W(\tau)$ and the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ as

$$W(\tau) \log a - a \sum_{j=1}^{W(\tau)} X_{(j)} + (n - W(\tau))[\log b - (a-b)\tau] - b \sum_{j=W(\tau)+1}^{n} \frac{X_{(j)}}{n!}$$

For $X_{(n-1)} \leq \tau < X_{(n)}$, the above expression is effectively

$$(n-1) \log a - a \sum_{j=1}^{n-1} X_{(j)} - a\tau + \log b - b(X_{(n)} - \tau)$$

If we take $\hat{b} = \frac{1}{X_{(n)} - \hat{\tau}} > 0$ and let $\hat{\tau}$ get close to $X_{(n)}$, $\log L_n$ will tend to ∞. Obviously, an unbounded function cannot have a maximum. MLE for θ does not exist.

Remarks

(i) In a situation such as in the above example, when a method of estimation like MLE cannpt be used, you should look for other methods of estimation (see a below section for some other methods). The point is this. When "considering" any method of estimation, you have to make sure that the method is "valid" in your problem under investigation, and not just "apply" some formulae!

(ii) *Why MLE is a popular method for estimation?*

Roughly speaking, for a class of "nice" models (called regular models), the MLE method provides consistent and asymptotically normal (asymptotically efficient) estimators.

b) Excess Mass Principle

The above MLE principle is useful for finding good estimators of *finitely dimensional* parameters of populations. For *infinitely dimensional parameters*, here is another principle for nonparametric estimation, especially for high dimensions.

Let X be a random vector with values in $\mathbb{R}^d$, having the unknown probability density function $f : \mathbb{R}^d \to \mathbb{R}^+$. Given a random sample $X_1, X_2, ..., X_n$ from X, we wish to estimate $f(x)$ at each given point $x \in \mathbb{R}^d$.

While there are methods such as kernel methods or orthogonal functions (which assume analytic properties of the unknown density), another approach, suggested by J.A. Hartigan (1987), could be used when qualitative information about the unknown density (such as its shape, geometric properties of its contour clusters) is available rather than analytic information.

For $\alpha > 0$, the $\alpha-$ level set of f is

$$A_\alpha(f) = A_\alpha = \{x \in \mathbb{R}^d : f(x) \geq \alpha\}$$

Since

$$f(x) = \int_0^\infty 1_{A_\alpha}(x) d\alpha$$

an estimator of $f(x)$, say, $f_n(x; X_1, X_2, ..., X_n)$ could be obtained, by plug-in, if we can estimate the *sets* A_α, for each $\alpha > 0$. But then, what is a plausible way to estimate A_α (of course, by some *random set estimator*) ? Again, by a plausible way, or a principle, we mean a way of estimation which could lead to a good estimator.

Let $dF(x) = f(x)d\mu(x)$, where $d\mu(x)$ denotes the Lebesgue measure on $\mathbb{R}^d$. Clearly, $(dF - \alpha\mu)(A_\alpha)$ is the "excess mass" of the set A_α at level α. Thus, we can consider the signed measure $dF - \alpha\mu = \varepsilon_\alpha$ on $\mathbb{R}^d$ with $\varepsilon_\alpha(A)$ being the excess mass of $A \in \mathcal{B}(\mathbb{R}^d)$ at level α. Writting

$$A = (A \cap A_\alpha) \cup (A \cap A_\alpha^c)$$

we see that, for any $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\varepsilon_\alpha(A) \leq \varepsilon_\alpha(A_\alpha)$$

i.e., the level set A_α has the largest excess mass at level α among all Borel sets. This suggests a way to estimate A_α by using the empirical counterpart of $dF - \alpha\mu$, namely

$$\varepsilon_{\alpha,n} = dF_n - \alpha\mu$$

where

$$dF_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

with δ_x being the Dirac measure at the point $x \in \mathbb{R}^d$.

The empirical excess mass, at level α, of A is $\varepsilon_{\alpha,n}(A) = (dF_n - \alpha\mu)(A)$.

It is expected that, as $n \to \infty$, $\varepsilon_{\alpha,n}(A)$ should converge to $\varepsilon_\alpha(A)$, so that maximizing $\varepsilon_{\alpha,n}(A)$ over $A \in \mathcal{C}$ (some subclass of Borel sets) could lead to good estimators of A_α.

Note that, while this principle is an optimization problem, it is an optimization of *set-functions* since $A \to (dF_n - \alpha\mu)(A)$ is a function whose argument A's are sets, and not a vector, or a function. *It needs special optimization technique!*

c) Maximum Entropy Principle

This is a principle for choosing a canonical probability distribution density function among a set of relevant ones. It started with Laplace (see E.T. Jaynes, "Where do we stand on maximum entropy?" in R. D. Levine and M. Tribus, Eds., *The Maximum Entropy Formalism*, MIT Press, Cambridge, Massachusetts, 1979, pp. 15-118). If your unknown distribution has support on a finite set or on a bounded interval, and no other information is available, then it makes sense to put the elements of the support on an equal footing, that is to endow it with a uniform distribution. This is referred to as *Laplace's insufficient reason principle.* Now (see below) the uniform distribution has the highest entropy among all densities on that support, the insufficient reason principle is equivalent to the principle of maximum entropy.

Motivated by the success of statistical mechanics, the principle of maximum entropy has been formalized as an inference procedure by Jaynes (Information theory and statistical mechanics, *Physical Review* (106), 620-630, (109), 171-127, 1957). Thus, if information about your distribution is that it is in a set $\mathcal{F}$ of densities, then the Maximum Entropy Principle postulates that you should seek the density whose entropy is

maximum over $\mathcal{F}$, i.e., solve

$$\max_{f \in \mathcal{F}} H(f) = \max\{-\int f(x) \log f(x) dx : f \in \mathcal{F}\}$$

Note that the class $\mathcal{F}$ is formed by constraints or evidence on possible densities.

Stochastic systems are expected to evolve into states with higher entropy as they approach equilibrium. The entropy of a probability density f is interpreted as a *measure of information* carried by f, where higher entropy means less information (more uncertainty or more lack of information). The main idea behind the maximum entropy principle is this. We should select a probability distribution which is consistent with our knowledge and introduce no unwarranted information. Any distribution (satisfying known constraints) which has smaller entropy will contain more information (less uncertainty) and hence says something stronger than what we know. The distribution with maximum entropy (satisfying our known constraints) is the one which should be least surprising in terms of the predictions it makes. The maximum entropy principle guides us to the best distribution which reflects our current knowledge and it tells us what to do if experimental data do not agree with predictions coming from our chosen distribution: look for previously unseen constraints and maximize entropy over all available constraints, including the new ones.

The Maximum Entropy Principle is useful in a variety of situations (e.g., in Econometrics) with several advantages such as:

(i) it incorporates as much (or as little) information as there is available,

(ii) it makes no assumption on a particular form of the joint distribution,

(iii) it can be applied to both numerical and qualitative random variables (as it only involves the distributions and not the "values" that the random "elements" take, for example, categorical variables such as those take "values" as "low, middle, high"),

(iv) it can take into account of any form of constraints, not only moments and linear correlations.

Hung T. Nguyen
Guest Editor