An Alternative to Maximum Likelihood Based on Spacings

Stanislav Anatolyev*
New Economic School, Moscow

Grigory Kosenok
New Economic School, Moscow

Problem and Motivation

Consider a scalar continuously distributed random variable x with probability density function (PDF) $f(x) \equiv f(x, \theta)$ and corresponding cumulative distribution function (CDF) $F(x) \equiv F(x, \theta)$ known up to a possibly multidimensional parameter $\theta \in \Theta$. Denote the true value of the parameter by θ_0. Suppose that a random sample $x_1, \cdots, x_n$ is available. To estimate θ_0, one usually uses the Maximum Likelihood (ML) estimator

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} Q_{ML}(\theta),$$

where

$$Q_{ML}(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \log f(x_{(i)}, \theta),$$

and $x_{(i)}$ is i^{th} order statistic. The ML procedure adjusts the shape of the density by tuning the parameter θ so that the product of density values at the sample observations is maximized. Suppose that the regularity conditions for consistency and asymptotic normality of the ML estimator hold; in particular, assume that the support of x does not depend on θ, and θ_0 is a well-separated point of maximum of $E[\log f(x, \theta)]$ (see, e.g., van der Vaart, 2000, ch. 5).

The following *Maximum Product of Spacings* (MPS) estimator has been encountered in the statistics literature but is not widespread among econometricians:

$$\hat{\theta}_{MPS} = \arg\max_{\theta \in \Theta} Q_{MPS}(\theta),$$

where

$$Q_{MPS}(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n+1} \log \left[F(x_{(i)}) - F(x_{(i-1)}) \right],$$

*Corresponding author. Address: New Economic School, Nakhimovsky pr., 47, room 1721, Moscow, 117418, Russia. E-mail: sanatoly@nes.ru

and $F(x_{(0)}) \equiv 0$, $F(x_{(n+1)}) \equiv 1$. The MPS criterion amounts to tuning the parameter θ so that the product of probabilities of a new observation falling between each two neighboring sample points in the set of order statistics is maximized (a similar interpretation was given by Titterington, 1985). Given that the sum of those probabilities equals one, at the optimum they are as close to each other as possible.

The MPS estimation was introduced independently by Cheng and Amin (1983) and Ranneby (1984). The advantages of MPS over ML were demonstrated when the boundedness of the objective function by construction leads to the consistency of MPS estimates when ML estimates may not exist or may not be unique, and in numerous examples when the density is positive only to the right of a shifted origin, it being one of unknown parameters (Cheng and Amin, 1983). In regular cases MPS estimates are consistent, asymptotically normal and efficient, which was derived in subsequent papers from the first principles (e.g. Shao and Hahn, 1999; Ghosh and Jammalamadaka, 2001).

However, in regular cases when both the ML and MPS estimators are well-defined, one can derive the asymptotic properties of the MPS by contrasting the two estimators to the first order as $n \to \infty$.

Solution and Discussion

We require that $f(x, \theta)$ be twice continuously differentiable in θ and continuous in x for all $\theta \in \Theta$. We will denote by x_L and x_U the minimal and maximal points of the support of x (either or both may be infinite). The support is supposed to be independent of the parameter θ and be connected, so that $f(x) > 0$ for any $x \in (x_L, x_U)$. Denote the true PDF by $f_0(x) \equiv f(x, \theta_0)$, and the true CDF by $F_0(x) \equiv F(x, \theta_0)$. We assume that $F(x, \theta)$ and $f(x, \theta)$ satisfy the following additional conditions.

Additional Assumption.

(A) $\lim\limits_{x \to x_L} F_0(x) \log^2 F(x) = \lim\limits_{x \to x_L} F_0(x) \log^2 f(x) = 0$,
$\lim\limits_{x \to x_U} (1 - F_0(x)) \log^2 (1 - F(x)) = \lim\limits_{x \to x_U} (1 - F_0(x)) \log^2 f(x) = 0$.

(B) For any $\theta \in \Theta$, a number of local optima of $f(x)$ is finite.

Part (A) imposes a sort of continuity on the tails as the parameter varies. Part (B) is not necessary, but it facilitates the proof. It is straightforward to verify that the Additional Assumption is satisfied for frequently used distributions. Our main result follows a helpful lemma.

Lemma 1 *If for some $g(x)$ we have $\lim\limits_{x \to x_L} F_0(x) g(x) = 0$, then $g(x_{(1)}) = o_p(n)$. Similarly, if for some $g(x)$ we have $\lim\limits_{x \to x_U} (1 - F_0(x)) g(x) = 0$, then $g(x_{(n)}) = o_p(n)$.*

Proof. The distribution of $F_0(x)$ is uniform on $[0,1]$ and hence (e.g., Feller, 1971)

$$\Pr\left\{nF_0(x_{(1)}) > t\right\} = \left(1 - \frac{t}{n}\right)^n \underset{n\to\infty}{\to} e^{-t}, \quad t \geq 0,$$

so that $nF_0(x_{(1)}) = O_p(1)$, but not $o_p(1)$, as $n \to \infty$. Observe that, as $\underset{n\to\infty}{\text{plim}} x_{(1)} = x_L$,

$$o_p(1) = F_0(x_{(1)})g(x_{(1)}) = nF_0(x_{(1)}) \cdot n^{-1}g(x_{(1)}) = O_p(1) \cdot n^{-1}g(x_{(1)}).$$

It follows that $g(x_{(1)}) = o_p(n)$. One handles the second implication in a similar manner. ■

Theorem 1 *Let the regularity conditions for consistency and asymptotic normality of the ML estimator $\hat{\theta}_{ML}$ hold. Let the Additional Assumption also hold. Then $\sqrt{n}(\hat{\theta}_{ML} - \hat{\theta}_{MPS}) = o_p(1)$.*

Proof. Denote $\Delta x_{(i)} \equiv x_{(i)} - x_{(i-1)}$. By the mean value theorem, for $i = 2, \cdots, n$,

$$F(x_{(i)}) - F(x_{(i-1)}) = \Delta x_{(i)} f(x^*_{(i)}),$$

where $x^*_{(i)} \in (x_{(i-1)}, x_{(i)})$. Then

$$\prod_{i=2}^{n} \left[F(x_{(i)}) - F(x_{(i-1)})\right] = \prod_{i=2}^{n} \Delta x_{(i)} \prod_{i=2}^{n} f(x^*_{(i)}) = \prod_{i=2}^{n} \Delta x_{(i)} \cdot \exp(\Lambda) \cdot \prod_{i=1}^{n} f(x_{(i)}),$$

where

$$\Lambda = \sum_{i=2}^{n} \log f(x^*_{(i)}) - \sum_{i=1}^{n} \log f(x_{(i)}).$$

Hence

$$Q_{MPS}(\theta) - Q_{ML}(\theta) = \frac{1}{n}\sum_{i=2}^{n} \log \Delta x_{(i)} + \frac{1}{n}\log F(x_{(1)}) + \frac{1}{n}\log(1 - F(x_{(n)})) + \frac{1}{n}\Lambda.$$

The first term is independent of θ and thus does not affect the optimizer. By Additional Assumption (A) and Lemma 1 appropriately applied with $g(x) = \log^2 F(x)$ and $g(x) = \log^2(1 - F(x))$, the second and third terms are $o_p(n^{-\frac{1}{2}})$.

Let us denote the K local optima of $f(x, \theta)$ by $\hat{x}_k$, $k = 1, \cdots, K$, and define in addition $\hat{x}_0 = x_L$ and $\hat{x}_{K+1} = x_U$. For every interval between adjacent points $\hat{x}_{k-1}$ and $\hat{x}_k$, $k = 1, \cdots, K+1$, denote by $x_{(i^l_k)}$ and $x_{(i^r_k)}$ the leftmost and rightmost order statistics that belong to the interval, and by $x^*_{(i_k)} \in (x_{(i^r_k)}, x_{(i^l_{k+1})})$, $k = 1, \cdots, K$ those $x^*_{(i)}$-points that do not belong to $\bigcup_{k=1}^{K+1}(x_{(i^l_k)}, x_{(i^r_k)})$. Then

$$\Lambda = \sum_{k=1}^{K} \log f(x^*_{(i_k)}) + \sum_{k=1}^{K+1} \Lambda_k, \quad \Lambda_k = \sum_{i=i^l_k+1}^{i^r_k} \log f(x^*_{(i)}) - \sum_{i=i^l_k}^{i^r_k} \log f(x_{(i)}).$$

The density $f(x,\theta)$ is monotone on any interval $[x_{(i_k^l)}, x_{(i_k^r)}]$. Suppose it is monotonically increasing. Then the greatest possible value taken by any $f(x^*_{(i)})$, $i_k^l + 1 \le i \le i_k^l$, is $f(x_{(i)})$. Thus Λ_k is bounded from above by $-\log f(x_{(i_k^l)})$. Analogously, Λ_k is bounded from below by $-\log f(x_{(i_k^r)})$. Therefore, $|\Lambda_k| \le |\log f(x_{(i_k^l)})| + |\log f(x_{(i_k^r)})|$. Similarly, the same inequality holds if $f(x,\theta)$ is monotonically decreasing on the interval considered. Summing over the intervals and rearranging we obtain

$$|\Lambda| \le |\log f(x_{(1)})| + |\log f(x_{(n)})| + \sum_{k=1}^{K} \left[|\log f(x^*_{(i_k)})| + |\log f(x_{(i_k^2)})| + |\log f(x_{(i_{k+1}^1)})| \right]$$

By Lemma 1 applied to $g(x) = \log^2 f(x)$, we have $|\log f(x_{(1)})| = o_p(n^{\frac{1}{2}})$ and $|\log f(x_{(n)})| = o_p(n^{\frac{1}{2}})$. The term with the summation is $O_p(1)$ because for all $k = 1, \cdots, K$

$$\operatorname*{plim}_{n\to\infty} |\log f(x^*_{(i_k)})| = \operatorname*{plim}_{n\to\infty} |\log f(x_{(i_k^r)})| = \operatorname*{plim}_{n\to\infty} |\log f(x_{(i_{k+1}^l)})| = |\log f(\hat{x}_k)|,$$

and K is finite.

To summarize, $Q_{MPS}(\theta) = Q_{ML}(\theta) + A + o_p(n^{-\frac{1}{2}})$, where A does not depend on θ. Therefore, $\hat{\theta}_{MPS} = \hat{\theta}_{ML} + o_p(n^{-\frac{1}{2}})$. ■

The asymptotic equivalence of the two estimators implies that the asymptotic or bootstrap inference about θ_0 on the basis of the MPS estimator may be carried out using the ML asymptotics. In small samples the unpleasant behavior of the ML estimator is well known. Using the alternative MPS criterion one may be able to improve the small sample performance of the ML estimator without changing the asymptotic distribution, as the following example attests.

Example 1 *The MPS estimator can be expected to be more efficient than the ML estimator in small samples when the density is skewed and/or the tails are heavy. Consider the exponential distribution* $f(x,\lambda) = \lambda e^{-\lambda x}\mathbb{I}[x \ge 0]$*, and let the true value* λ_0 *be* 1*. The mean squared error of* λ_{MPS} *obtained from* 10,000 *simulations is 65.4%, 73.1% and 85.2% that of* λ_{ML} *for sample sizes of* 6, 10 *and* 20*, respectively. Very similar comparisons emerge for the Pareto distribution* $f(x) = \lambda x^{-(\lambda+1)}\mathbb{I}[x \ge 1]$ *with the true value* λ_0 *being 1.*

Of course, the efficiency property is not universal: the MPS and ML estimators are equally efficient when a mean of a normal distribution with known variance is estimated, while when a standard deviation of a centered normal distribution is estimated, the mean squared error of the MPS estimator exceeds that of the ML estimator by the factor that ranges from 2.7 to 1.4 as the sample size increases from 6 to 50.

As far as extensions of the MPS idea are concerned, Lind (1994) provides a rationale for the MPS in the information theory, Ekström (1997) suggests use of "high order" spacings, Ghosh and Jammalamadaka (2001) expand the class to contain estimators with objective functions based on various divergence criteria (interestingly, the MPS estimator is the only one asymptotically efficient in the extended class), and Ranneby, Jammalamadaka and Teterukovskiy (2004) propose a generalization to multivariate situations.

Acknowledgements

We thank the coeditor Paolo Paruolo for his patience, and an anonymous referee for providing references to the statistics literature.

References

Cheng, R.C.H. & N.A.K. Amin (1983) Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society Series B* 45, 394–403.

Ekström, M. (1997) Generalized maximum spacing estimators. Manuscript, Umeå University.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications.* John Wiley & Sons, New York.

Ghosh, K. & S.R. Jammalamadaka (2001) A general estimation method using spacings. *Journal of Statistical Planning and Inference* 93, 71–82.

Lind, N.C. (1994) Information theory and maximum product of spacings estimation. *Journal of the Royal Statistical Society Series B* 56, 341–343.

Ranneby, B. (1984) The maximum spacing method: An estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics* 11, 93–112.

Ranneby, B., Jammalamadaka, S.R. & A. Teterukovskiy (2004) The maximum spacing estimate for multivariate observations. Manuscript, Swedish University of Agricultural Sciences.

Shao, Y. & M. Hahn (1999) Strong consistency of the maximum product of spacings estimates with applications in nonparametrics and in estimation of unimodal densities. *Annals of the Institute of Statistical Mathematics* 51, 31–49.

Titterington, D.M. (1985) Comment on "Estimating parameters in continuous univariate distributions". *Journal of the Royal Statistical Society Series B* 47, 115–116.

van der Vaart, A.W. (2000) *Asymptotic Statistics.* Cambridge University Press, Cambridge.