Power Law Analysis for Total Energy Data

Xufei Wang

February 28, 2017

Parametric MPS

Let $X_1, \ldots, X_n$ be independent observation from a continuous univariate distribution F_{θ_0} belonging to $\{F_\theta : \theta \in \Theta\}$. Consider the problem of estimating the unknown true θ_0. Applying the cdf F_θ to the order statistics $X_{1,n} \leq \ldots \leq X_{n,n}$ yields:

$$0 = F_\theta(X_{0,n}) \leq F_\theta(X_{1,n}) \leq \ldots \leq F_\theta(X_{n,n}) \leq F_\theta(X_{n+1}, n) = 1.$$

The maximum product of spacings methods chooses $\hat{\theta}_n$, which maximizes the product of spacing, i.e.,

$$\hat{\theta}_n = \mathrm{argmax}_\theta \prod_{i=1}^{n+1} \left[F_\theta(X_{i,n}) - F_\theta(X_{i-1,n})\right].$$

Note that the spacings sum to 1.

MPS and MLE

Note that when F_θ has a density f_θ, the log likelihood can be written as

$$\sum_{i=1}^{n} \log f_\theta(X_{i,n});$$

while the log product of spacings can be approximated by

$$\sum_{i=1}^{n+1} \log f_\theta(X_{i,n}) + \sum_{i=1}^{n+1} \log(X_{i,n} - X_{i-1,n}).$$

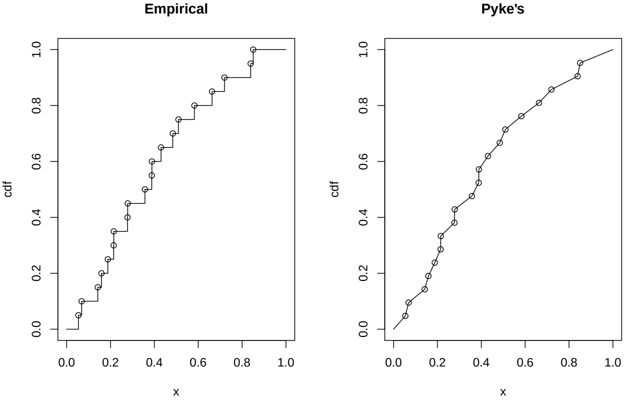
Since $\sum_{i=1}^{n+1} \log(X_{i,n} - X_{i-1,n})$ is a constant, this suggests that the two methods should lead to similar estimates when the MLE method works.

On the other hand, the product of spacings is always bounded and hence is more stable then the likelihood functions. The MPS method can generate asymptotically optimal estimates even when ML breaks down due to unbounded likelihood functions. Shao (1999) proves that under certain conditions, with probability one an MPS estimator exists for all sufficiently large n, and any MPS estimator is consistent.

Nonparametric MPS for a concave distribution function

Suppose that the underlying cdf F_0 has a non-increasing density on $[0, 1]$, Grenander (1956) applied the ML method and obtained the nonparametric MLE of F_0: the lease concave majorant (LCM) of the empirical cdf, F_n. This estimator is also asymptotically optimal in the minimax sense. While the empirical cdf F_n is not continuous, Pyke's modified empirical π_n is a continuous cdf which puts equal mass $\frac{1}{n+1}$ on every spacing and is uniformly distributed on each finite spacing. π_n is a MPS estimator of F_0. Shao (2001) proves that the nonparametric MPS estimator for a concave cdf is asymptotically minimax has the following simple form: The MPS estimator for a concave cdf with known finite support is the lease concave majorant (LCM) of Pyke's modified empirical distribution π_n.

Empirical cdf and Pyke's modified empirical cdf



Power law data

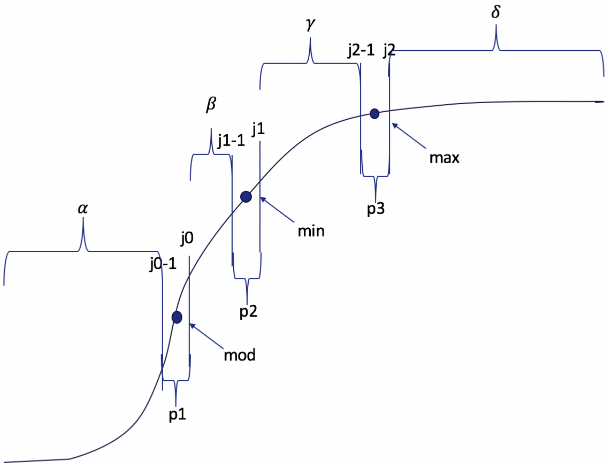
Suppose a continuous random variable X follows power law distribution, then for x large enough, we have $f_X(x) \propto x^{-\alpha}$. Suppose $Y = \log X$, then for y large enough, $f_Y(y) \propto e^{-(\alpha-1)y}$.

Total energy data

Let $X_1, \ldots, X_n$ be the total energy data. The distribution of X_i is a unimodal distribution and suppose the mode is x_{mod}. By physics knowledge, we know that for $X \in (x_{min}, x_{max})$, the distribution follows power law distribution. Let $Y_i = \log X_i$, $y_{mod} = \log x_{mod}$, $y_{min} = \log x_{min}$ and $y_{max} = \log x_{max}$. Then Y_i also follows a unimodal distribution with the mode y_{mod} and when $Y \in (y_{min}, y_{max})$, $f_y \propto e^{-(\alpha-1)y}$. Our interest in this problem is to get an estimate for y_{min}, y_{max} and α, especially for y_{max}. For simplicity, we suppose that Y is bounded.

MPS for total energy data

Suppose $Y_{1,n} \leq \ldots \leq Y_{n,n}$. Suppose the mode is on the j_0th spacing, y_{min} is on the $j_1$the spacing and y_{max} is on the j_2th spacing. Then F_0 is convex on $(-\infty, Y_{j_0-1,n}]$, concave on $[Y_{j_0,n}, Y_{j_1-1,n}]$ and concave on $[Y_{j_2,n}, \infty)$. Suppose $\alpha_F = F(Y_{j_0-1,n})$, $\beta_F = F(Y_{j_1-1,n}) - F(Y_{j_0,n})$, $\gamma_F = F(Y_{j_2-1,n}) - F(Y_{j_1,n})$ and $\delta_F = 1 - F(Y_{j_2,n})$. Besides, let $p_0 = F(Y_{j_0,n}) - F(Y_{j_0-1,n})$, $p_1 = F(Y_{j_1,n}) - F(Y_{j_1-1,n})$ and $p_2 = F(Y_{j_2,n}) - F(Y_{j_2-1,n})$.


δ
γ
j2-1
j2
β
j1
j1-1
max
p3
min
α
j0
j0-1
p2
mod
p1

MPS for total energy data

Then $\alpha_F + p_0 + \beta_F + p_1 + \gamma_F + p_2 + \delta_F = 1$. The the product of spacing is

$$
\begin{aligned}
& \prod_{i=1}^{n+1} [F(Y_{i,n}) - F(Y_{i-1,n})] \\
= \;& p_1 p_2 p_3 \prod_{i=1}^{j_0-1} [F(Y_{i,n}) - F(Y_{i-1,n})] \prod_{i=j_0+1}^{j_1-1} [F(Y_{i,n}) - F(Y_{i-1,n})] \\
& \prod_{i=j_1+1}^{j_2-1} [F(Y_{i,n}) - F(Y_{i-1,n})] \prod_{i=j_2+1}^{n+1} [F(Y_{i,n}) - F(Y_{i-1,n})] \\
= \;& p_1 p_2 p_3 \alpha_F^{j_0-1} \beta_F^{j_1-j_0-1} \gamma_F^{j_2-j_1-1} \delta_F^{n-j_2+1} A_{j_0} B_{j_0,j_1} C_{j_1,j_2} D_{j_2}
\end{aligned}
$$

The maximum of B_{j_0,j_1} and D_{j_2} is the LCM of the Pyke's empirical distribution and the maximum of A_{j_0} is the greatest concave minorant of the Pyke's empirical distribution.

MPS for total energy data

For C_{j_1,j_2}, the maximum is obtained with

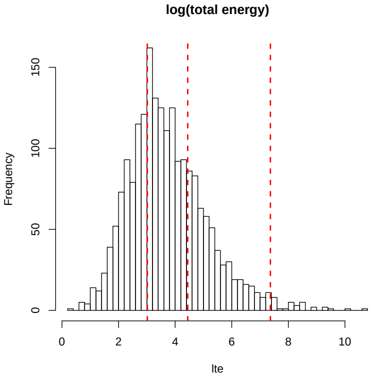
$$\begin{aligned}
\hat{\alpha} &= \operatorname{argmax}[e^{-(\alpha-1)Y_{j_1,n}} - e^{-(\alpha-1)Y_{j_2-1,n}}]^{j_2-j_1-1} \\
&\quad \prod_{i=j_1+1}^{j_2-1} [e^{-(\alpha-1)Y_{i-1,n}} - e^{-(\alpha-1)Y_{i,n}}] \\
&= \operatorname{argmax}[1 - e^{-(\alpha-1)(Y_{j_2-1,n}-Y_{j_1,n})}]^{j_2-j_1-1} \\
&\quad \prod_{i=j_1+1}^{j_2-1} [e^{-(\alpha-1)(Y_{i-1,n}-Y_{j_1,n})} - e^{-(\alpha-1)(Y_{i,n}-Y_{j_1,n})}]
\end{aligned}$$

The maximum of the product of spacing does not depend on A_{j_0}, B_{j_0,j_1}, C_{j_1,j_2} and D_{j_2}. The maximum is obtained when

$$p_1 = p_2 = p_3 = \frac{1}{n},\ \alpha_F = \frac{j_0 - 1}{n+1},$$
$$\beta_F = \frac{j_1 - j_0 - 1}{n+1},\ \gamma_F = \frac{j_2 - j_1 - 1}{n+1},\ \delta_F = \frac{n - j_2 + 1}{n+1}.$$

MPS for real data

A rough estimation shows that $y_{mod} \in (2.8, 3.5)$, $y_{min} \leq 5$ and $y_{max} \geq 7$.



MPS for real data

We compute the every possible combination of (j_0, j_1, j_2) and report the maximum result as $y_{mod} = 3.01$, $y_{min} = 4.44$ and $y_{max} = 7.37$ with $\hat{\alpha} = 1.8624$. If y_{mod}, y_{min} and y_{max} are known, the consistency and minimaxity of the MPS estimate is guaranteed. However, when they are unknown, the consistency and minimaxity is an open problem.