The influence of small sample sizes on the determination of power laws

How large does your sample need to be before you can publish a power law?

E. D'Huys, D.B. Seaton, D. Berghmans

STCE - SIDC, Royal Observatory of Belgium

The frequency distributions of different solar parameters show a power law behavior, often interpreted based on the concept of self-organized criticality (SOC, Bak et al. 1988). In the corona, Lu & Hamilton (1991) argued that a solar flare can be interpreted as an avalanche of many small reconnection events, resulting in a power law distribution for the flare occurrence. This power law implies that flares are scale-invariant: flares of all sizes are the result of the same physical process and their strength is determined by the number of elementary reconnection events involved. Robbrecht et al. (2009) studied CME widths and, similarly, found a power law behavior over a large range of angular widths.

We studied the scale invariance of the widths of CMEs that were not associated with EUV signatures in the low corona ("stealth" CMEs) and observed an important influence of the sample size on the derived slope (on a logarithmic scale). When selecting only a small sample of CMEs, the resulting slope is notably smaller than the one derived for a large number of CMEs by Robbrecht et al. This flatter distribution is not surprising as only a small random selection of many CMEs is made, and including a wide CME in such a small sample decreases the slope significantly. In reality the CME angular width distribution is dominated by narrow events, which becomes clear when more detections are taken into account. Based on an artificial distribution, we investigate the influence of the sample size on the derived power law.

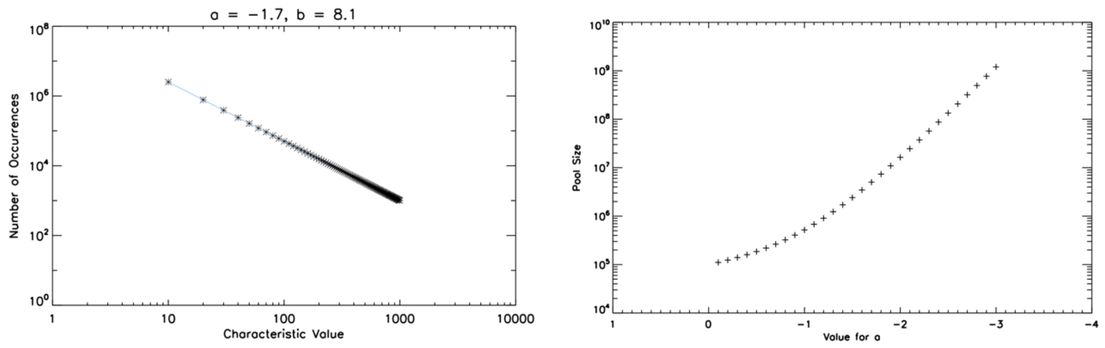

Fig. 1 An artificial power law distribution with scaling parameter a = -1.7 (left). This exponent constitutes the slope of the straight line that becomes apparent when the distribution is plotted in log-log space ($\log(y) = a * \log(x) + b$). The parameter b is chosen to ensure a sufficiently large number of elements in the distribution. This pool size is shown on the right for different values of the scaling parameter a.

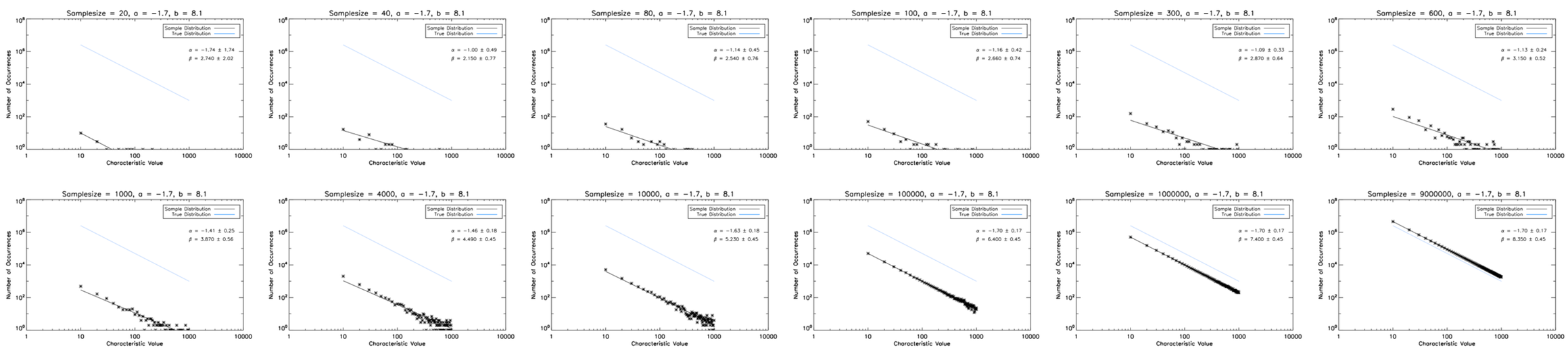

Fig. 2 Best estimates of the power law parameters with samples of different sizes (between 20 and 9×10^6) using a least-squares line fitting method. The blue line indicates the true power law distribution. For small sample sizes, the fit is poor and dominated by the small characteristic values. As the sample size increases, more high characteristic values are included, and the fit improves. However, there is a problem of under-sampling on the right-hand-side of the slope where there are only a few elements per bin. This is resolved when an even larger sample size is chosen.

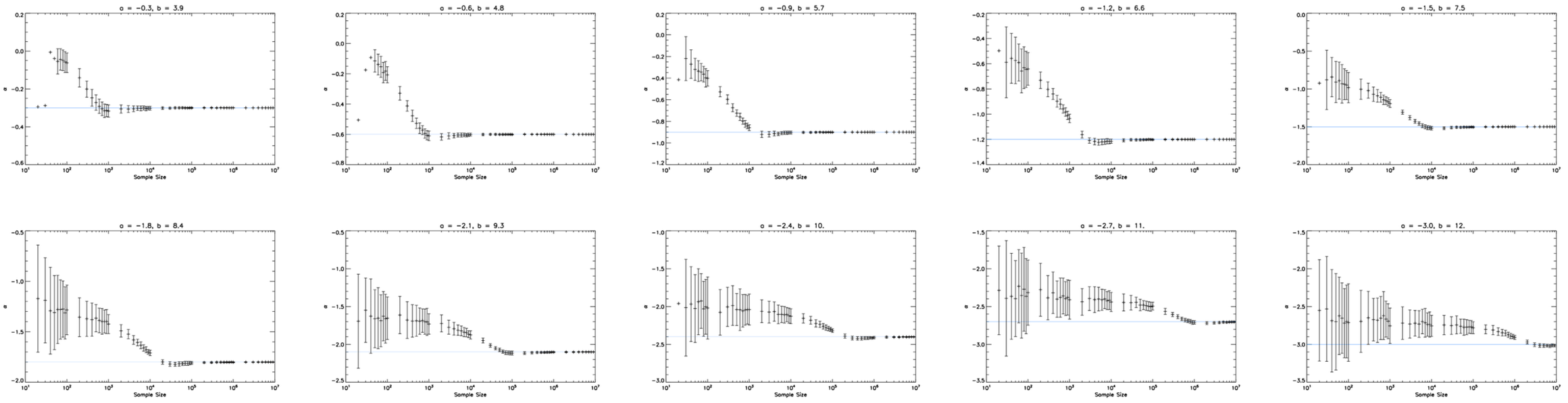

Fig. 3 Convergence of the estimated scaling parameter **α** towards its true value **a** as a function of the sample size and for different values of **a**. To eliminate the strong fluctuations for small sample sizes, we show the average results over 100 random and independent samples. The error bars indicate the standard deviation of **α** for the 100 samples. It is immediately evident that large sample sizes are needed to reliably estimate the power law slope with this least-squares line fitting method (at least 10^3 elements). Additionally, the needed sample size increases as the slope becomes steeper. This is due to the sensitivity of steep power laws to the random selection of large values: for a flatter distribution it is less important which elements are randomly selected from the distribution as the behavior over the complete x-range is the same; on the other hand, for a steep slope sampling elements in the higher x-range will flatten the slope significantly. A larger sample is less sensitive to this effect.
Note that for steeper slopes (second row) the estimated scaling parameter **α** reaches a plateau value before a critical increase in sample size ensures sufficient data points in the highest bins, and the true value of **a** is reached.

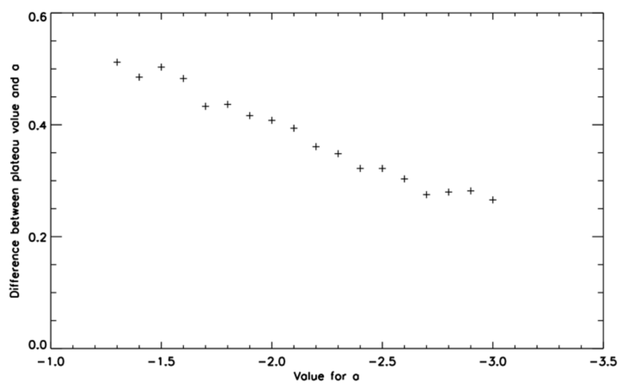

Fig. 4 The difference between the plateau values reached in Fig. 3 and the true scaling parameter value **a**. This difference decreases to ~0.3 for steep power laws. Possibly these plateau values can be used to estimate the true parameter value in case only a limited sample can be acquired.

Conclusions

To reliably estimate the scaling parameter for a power law distribution using a least-squares line fitting method, a large sample size is needed.

This required sample size increases when the scaling parameter is larger (steeper slope).

The estimate of the scaling parameter can be improved by changing the bin size in case of under-sampling to ensure a sufficiently large number of elements in each bin, or by using logarithmic bins.

Other techniques are more appropriate to estimate the scaling parameter than a least-squares line fitting method. Most advertised is the Mean Likelihood Estimator (MLE), as described for example by Clauset et al. (2009).

Our results indicate that, depending on the sample size, the slope of previously published power laws might have been underestimated by 0.3 - 0.5. This means that the widely confirmed power law in the range -1.6 to -1.7 that is found for the soft X-ray flare peak flux (Aschwanden 2012) might actually be as steep as -2.0, the critical value to explain the heating of the corona by nanoflares (Hudson 1991)