Draft - Alternative Distance Function

Galen Wilkerson

March 27, 2013

Abstract

Does it make sense to use a linear distance function (subtraction) when evaluating the distance of two (very) non-linear functions such as power laws? Quick writeup of a problem with the KS statistic and an alternative possible way to find the distance between two power laws.

Linear Distance

In their article, Clauset, Shalizi, and Newman propose to use the Kolmogorov-Smirnov Statistic when evaluating the distance between observed data and the best-fit function.

These authors cite *Numerical Recipes in C* as their authority that, among distance metrics for non-normal data, the "commonest is the Kolmogorov-Smirnov or KS statistic, which is simply the maximum distance between the CDFs of the data and the fitted model":

$$D = \max_{x \geq x_{min}} |S(x) - P(x)|,$$

where $S(x)$ is the CDF of the observations and $P(x)$ is the CDF for the best-fit power law in the region $x \geq x_{min}$.

Clauset et al. use this KS statistic for finding x_{min} and for finding the p-value of the fit, so it is rather important. However, we might ask, is it appropriate to use a linear distance function between two non-linear functions? After all, in a power-law, the linear distance may be very large as x approaches zero. A linear distance is not given equal weight over all x values, and does not focus on the parameter of importance - the exponent.

Let there be two power law functions:

$$S(x) = \int_0^\infty Cx^{-\alpha}dx, \qquad P(x) = \int_0^\infty Dx^{-\beta}dx,$$

defined as above (observed and best-fit, respectively).

We then obtain:

$$S(x) = \frac{1}{1-\alpha}x^{1-\alpha}, \qquad P(x) = \frac{1}{1-\beta}x^{1-\beta},$$

giving the KS statistic of these two power-law functions,

$$D = \max_{x \geq x_{min}} \left| \frac{1}{1-\alpha}x^{1-\alpha} - \frac{1}{1-\beta}x^{1-\beta} \right|.$$

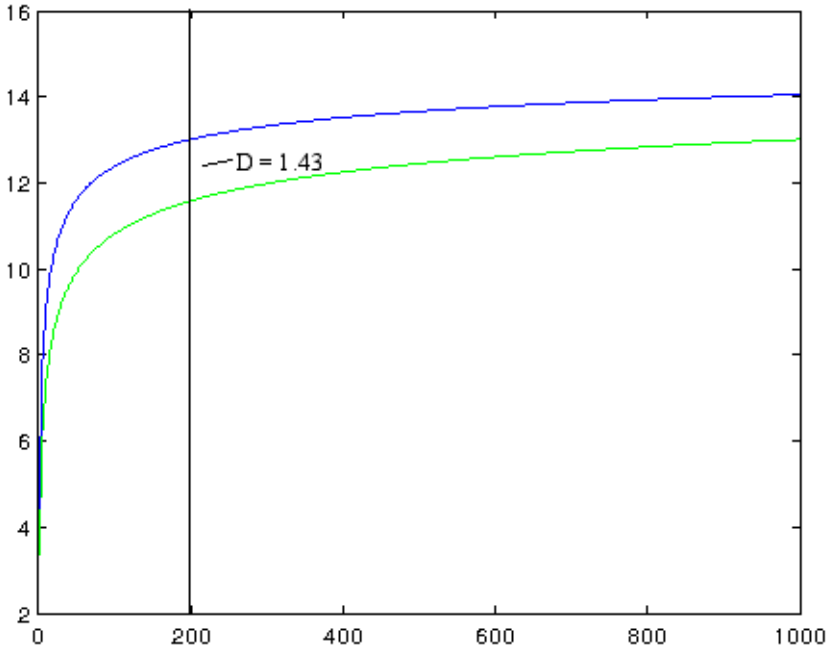

Figure 1: Here, we arbitrarily choose $S(x) = 3x^{-1.2}$ (green) and $P(x) = 4x^{-1.3}$ (blue). In this case, if we let $x_{min} = 200$, the KS statistic finds the maximum distance between the CDFs to be at x_{min} itself, rather than in the tail.

However, as we see in Figure , the two functions may differ near x_{min} or in the tail [this is claimed by Clauset]. This seems to be a function of the relationship between the coefficients and the exponents, but is not as rigorous as we could hope, since there are other tools at our disposal.

Non-Linear Distance

It seems perhaps more reasonable to take the logs of the two power laws, then find the linear distance, since we know that:

$$\log Cx^{-\alpha} = \log C - \alpha \log x$$

Thus, if $p(x) = Cx^{-\alpha}$ and $q(x) = Dx^{-\beta}$, we can find the 'distance' between these two functions, (in particular their exponents) by

$$D = \max_{x \geq x_{min}} |\log p(x) - \log q(x)|.$$

This has the advantage of truly focusing on the tail of the distributions, rather than the region near x_{min}.

Perhaps it is desirable to use the log of the CDF instead of the PDF, but first we have to understand why the CDF is used in the KS statistic. (I think it is because the CDF accumulates all of the difference between the two functions, so the authors Clauset et al. think this will demonstrate a difference in the tail of the two distributions in question.)

This method can be tested similarly to the tests performed in the article by Clauset et al.