The poweRlaw package: Comparing distributions

Colin S. Gillespie

Last updated: January 10, 2019

The poweRlaw package provides an easy to use interface for fitting and visualising heavy tailed distributions, including power-laws. This vignette provides examples of comparing competing distributions.

1 Comparing distributions

This short vignette aims to provide some guidance when comparing distributions using Vuong’s test statistic. The hypothesis being tested is

\[ H_0 : \text{Both distributions are equally far from the true distribution} \]

and

\[ H_1 : \text{One of the test distributions is closer to the true distribution.} \]

To perform this test we use the `compare_distributions` function\(^1\) and examine the `p_two_sided` value.

2 Example: Simulated data 1

First let’s generate some data from a power-law distribution

```r
library("poweRlaw")
set.seed(1)
x = rpldis(1000, xmin=2, alpha=3)
```

and fit a discrete power-law distribution

```r
m1 = displ$new(x)
m1$setPars(estimate_pars(m1))
```

The estimated values of \( x_{\text{min}} \) and \( \alpha \) are 2 and 2.97, respectively. As an alternative distribution, we will fit a discrete Poisson distribution\(^2\)

\(^1\)The `compare_distributions` function also returns a one sided \( p \)-value. Essentially, the one sided \( p \)-value is testing whether the first model is better than the second, i.e. a one sided test.

\(^2\)When comparing distributions, each model must have the same \( x_{\text{min}} \) value. In this example, both models have \( x_{\text{min}} = 2 \).
Figure 1: Plot of the simulated data CDF, with power law and Poisson lines of best fit.

\begin{verbatim}
m2 =dispois$new(x)
m2$setPars(estimate_pars(m2))

plot(m2, ylab="CDF")
lines(m1)
lines(m2, col=2, lty=2)

suggests that the power-law model gives a better fit (figure 1). Investigating this formally

comp = compare_distributions(m1, m2)
comp$p_two_sided

## [1] 0.05141726

means we can reject \( H_0 \) since \( p = 0.05142 \) and conclude that one model is closer to the true distribution.

**One or two-sided \( p \)-value**

The two-sided \( p \)-value does not depend on the order of the model comparison

\begin{verbatim}
compare_distributions(m1, m2)$p_two_sided

## [1] 0.05141726

compare_distributions(m2, m1)$p_two_sided

## [1] 0.05141726
\end{verbatim}

However, the one-sided \( p \)-value is order dependent
Comparing distributions

```r
## We only care if m1 is better than m2
## m1 is clearly better
compare_distributions(m1, m2)$p_one_sided

## [1] 0.02570863

## m2 isn't better than m1
compare_distributions(m2, m1)$p_one_sided

## [1] 0.9742914
```

3 Example: Moby Dick data set

This time we will look at the Moby Dick data set

```r
data("moby")
```

Again we fit a power law

```r
m1 = displ$new(moby)
m1$setXmin(estimate_xmin(m1))
```

and a log-normal model\(^3\)

```r
m2 = dislnorm$new(moby)
m2$setXmin(m1$getXmin())
m2$setPars(estimate_pars(m2))
```

Plotting the CDFs

```r
plot(m2, ylab="CDF")
lines(m1)
lines(m2, col=2, lty=2)
```

suggests that both models perform equally well (figure 2). The formal hypothesis test

```r
comp = compare_distributions(m1, m2)
gives a p-value and test statistic of

```r
comp$p_two_sided

## [1] 0.6773366
```

```r
comp$test_statistic

## [1] 0.4161003
```

which means we can not reject \(H_0\). The \(p\)-value and test statistic are similar to the values found in table 6.3 of Clauset et al. (2009).

\(^3\)In order to compare distributions, \(x_{\text{min}}\) must be equal for both distributions.
Comparing distributions

Figure 2: The Moby Dick data set with power law and log normal lines of best fit.

References