The poweRlaw package: Comparing distributions

Colin S. Gillespie

Last updated: May 27, 2014

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 look at the p_two_sided value.

2 *Example: Simulated data*

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

```
set.seed(1)
x = rpldis(10000, xmin=2, alpha=2.1)
```

and fit a discrete power-law distribution

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

The estimated values of $x_{\min}$ and α are 2 and 2.0947, respectively. As an alternative distribution, we will fit a discrete log-normal distribution

```
m2 = dislnorm$new(x)
m2$setPars(estimate_pars(m2))
```

When comparing distributions, each model must have the same $x_{\min}$ value. In this example, both models have $x_{\min} = 2$.

Plotting both models

```
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

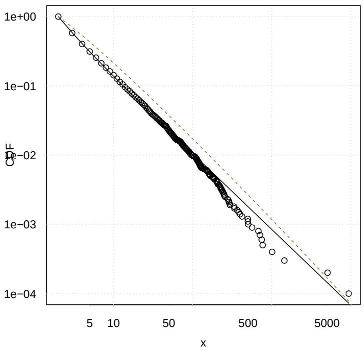

Figure 1: Plot of the simulated data CDF, with power law and log normal lines of best fit.

[1] The compare_distributions function also returns a one sided p-value. Essentially, the one side p-value is testing whether the first model is better than the second, i.e. a **one** sided test.

```
comp = compare_distributions(m1, m2)
comp$p_two_sided
## [1] 0.005119
```

means we can reject H_0 since $p = 0.0051$ and conclude that one model is closer to the true distribution.

3 *Example: Moby Dick data set*

This time we will look at the Moby Dick data set

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

Again we fit a power law

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

and a log-normal model[2]

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

Plotting the CDFs

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

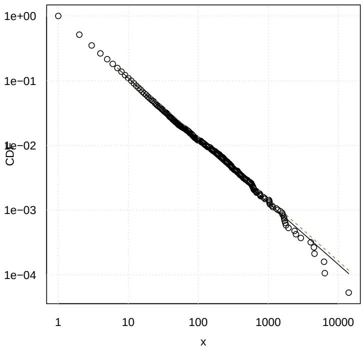

Figure 2: The Moby Dick data set with power law and log normal lines of best fit.

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

```
comp = compare_distributions(m1, m2)
```

gives a p-value and test statistic of

```
comp$p_two_sided
## [1] 0.6824

comp$test_statistic
## [1] 0.4092
```

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.[3].

[2] In order to compare distributions, $x_{\min}$ must be equal for both distributions.

[3] A.~Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009

References

A.~Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

Session Info

```
print(sessionInfo(), locale = FALSE)
## R version 3.1.0 (2014-04-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## attached base packages:
## [1] stats     graphics  grDevices utils
## [5] datasets  methods   base
##
## other attached packages:
## [1] poweRlaw_0.20.3 knitr_1.6
##
## loaded via a namespace (and not attached):
## [1] VGAM_0.9-3     evaluate_0.5.5 formatR_0.10
## [4] highr_0.3      parallel_3.1.0 stats4_3.1.0
## [7] stringr_0.6.2  tools_3.1.0
```