Tutorials on AstroStatistics and R, by Eric Feigelson

Jan 29 & Jan 31, 2014, Phillips Auditorium, CfA

`http://hea-www.harvard.edu/AstroStat/Tutorial2014/`

Thank you all for your participation! I intend to do 3 things over the next week:
1. ~~get the slides from Eric and put them online (but accessible only from CfA IPs)~~ [done]
2. ~~put the recorded videos on YouTube (accessible only through your sign-in)~~ [done]
3. ~~Freeze this doc and put a copy online (so if you want to add/edit stuff with commentary or questions or something else, please do so by Monday)~~ [done]

I will send out another email to all registrants after all 3 things are done.

If you have feedback on how this webcast/scratch space worked, and/or if you can figure out ways to improve it, you know where I am!

-Vinay

Before the Tutorials begin, please download and install R on your laptops. The current version is v3.0.2, released 2013/09/25.

- To download R, go to http://www.r-project.org
  - There are binaries for MacOS and Windows, and a short installation process for different versions of Linux
- Some R FAQs are at http://cran.r-project.org/faqs.html
- R Manuals are at http://cran.r-project.org/manuals.html
- For Mac OS X, the command-line version gets installed at `/usr/bin/R`
- For people who are familiar with python, see below (thx Scott) for a description of how to install `pyR`.

R scripts that will form the foundation of the tutorials are at http://www2.astro.psu.edu/users/edf/CfA_Jan2014/ -- download them now (because who knows how good the internet connection will be under the expected heavy load) and try them out! These require you to install some packages and use some data. The packages are

`cobs, quantreg, MatrixModels, np, nlstools, fpc, class, e1071, rgl, scatterplot3d, spdep, gstat, spatstat, geoR`

and the datasets are:

- `http://astrostatistics.psu.edu/MSMA/datasets/SDSS_QSO.dat`

- http://astrostatistics.psu.edu/MSMA/datasets/NGC4472_profile.dat
- http://astrostatistics.psu.edu/MSMA/datasets/NGC4406_profile.dat
- http://astrostatistics.psu.edu/MSMA/datasets/NGC4551_profile.dat
- http://astrostatistics.psu.edu/MSMA/datasets/COMBO17_lowz.dat
- http://astrostatistics.psu.edu/MSMA/datasets/Shapley_galaxy.dat
- http://astrostatistics.psu.edu/MSMA/datasets/SDSS_test.csv
- http://astrostatistics.psu.edu/MSMA/datasets/SDSS_stars.csv
- http://astrostatistics.psu.edu/MSMA/datasets/SDSS_wd.csv

The psu.edu site seems to be having bandwidth problems. I have downloaded the full (as far as I can tell) versions to our Tutorial website. Just replace all the "astrostatistics.psu.edu/MSMA/datasets" with "hea-www.harvard.edu/AstroStat/Tutorial2014" in the above links.

Eric will be sitting in Room C311-D during his visit.

The tutorials will be webcast. We encourage people to check-in from the comfort of their offices. Interactivity will be made possible through this page (though the latency will not compare well with being at the Auditorium in person). We will monitor this page during the tutorials and convey your questions and comments to Eric.

Type comments and questions below this line. (Remember to identify yourself!)

Hi vinay,

Just to be sure. So are we supposed to stay in the office? That's great.
In that case, I would use R in the linux system, which appears to be in a wide range of different versions, from 2.15.1 to 3.0.2.

[VK] No, you are welcome to choose either option. I think personally it will be better if those who can make it do come to Phillips. The webcast thing is an experiment. It is a trade-off between latency and elbow room. We will see how it goes!

[VK] Download v3.0.2. That is the latest one, and there are some incompatibilities between versions 2 and 3.

OK, I will do that way. BTW, this is Dong-Woo.

Hi Folks - non- science comment from Michelle Henson: **Please note that this is a bring your own coffee/tea/soda event.** We are working to provide light refreshments in the

morning and afternoon.

From SJW: ok guys I was able to get R installed under python on the mac
I first installed the R 3.0.2 package from the R site (R-3.0.2.pkg).
I then ran easy_install from the command line:
...
lothlorien:~>easy_install rpy2
Searching for rpy2
Reading http://pypi.python.org/simple/rpy2/
Best match: rpy2 2.3.9
….
zip_safe flag not set; analyzing archive contents...
rpy2.rinterface.tests.test_EmbeddedR: module references __path__
Adding rpy2 2.3.9 to easy-install.pth file

Installed
/home/swolk/anaconda/lib/python2.7/site-packages/rpy2-2.3.9-py2.7-macosx-10.5-x86_64
.egg
Processing dependencies for rpy2
Finished processing dependencies for rpy2
lothlorien:~>pylab
Python 2.7.5 |Anaconda 1.8.0 (x86_64)| (default, Oct 24 2013, 07:02:20)
Type "copyright", "credits" or "license" for more information.

IPython 1.1.0 -- An enhanced Interactive Python.
?   -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help   -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
Using matplotlib backend: MacOSX

In [2]: import rpy2

In [3]: import rpy2.tests
Error in loadNamespace(name) : there is no package called ‘ggplot2’

In [4]: import unittest

In [5]:

[Doug Burke here] Looking through the scripts at http://www2.astro.psu.edu/users/edf/CfA_Jan2014/ I see there's a numner of extra packages it is suggested we install. It's probably better if people try this before hand - do a 'grep install *txt' to find out all the install.packages calls needed, although I did find a couple (scatterplot3d and rgl spring to mind) in the Ubuntu repository.

Also, I think quotes are needed around cobs in the install.packages(cobs) statement in the 3_Density estimation_R.txt tutorial.

This is Frank. Just checking how things are working. I downloaded the R scripts and started running through them. Is it possible to 'import' them into an R session, or should I just cut-and-paste?

[VK] I think they are designed for cut-n-paste, so a step-by-step working through.

[Doug again] Similarly there are a bunch of datasets referenced in the tutorials that are probably worth downloading before tomorrow - particularly SDSS_QSO.dat.;

[VK] Good point! I will add the URLs for the .dat files in the preamble above.

[TA] Python script to download all the data files. I had trouble with the server giving up midway on the first (8.5 Mb) file. This requires the `requests` package, which is available by default in Anaconda (or "easy_install requests").

```
import os
import requests

urls = """
http://astrostatistics.psu.edu/MSMA/datasets/SDSS_QSO.dat
http://astrostatistics.psu.edu/MSMA/datasets/NGC4472_profile.dat
http://astrostatistics.psu.edu/MSMA/datasets/NGC4406_profile.dat
http://astrostatistics.psu.edu/MSMA/datasets/NGC4551_profile.dat
http://astrostatistics.psu.edu/MSMA/datasets/COMBO17_lowz.dat
http://astrostatistics.psu.edu/MSMA/datasets/Shapley_galaxy.dat
http://astrostatistics.psu.edu/MSMA/datasets/SDSS_test.csv
http://astrostatistics.psu.edu/MSMA/datasets/SDSS_stars.csv
```

```
http://astrostatistics.psu.edu/MSMA/datasets/SDSS_wd.csv
"""
urls = urls.strip().splitlines()

for url in urls:
    filename = os.path.basename(url)
    print('Getting {!r}'.format(url))
    fetch = requests.get(url)
    with open(filename, 'w') as fh:
        fh.write(fetch.text)
```

[VK] Here's another way to download all at one shot:

```
#!/bin/tcsh -f

set fils = "SDSS_QSO.dat NGC4472_profile.dat NGC4406_profile.dat NGC4551_profile.dat COMBO17_lowz.dat
Shapley_galaxy.dat SDSS_test.csv SDSS_stars.csv SDSS_wd.csv" # single line

foreach fil ( $fils )
  if ( ! -f $fil ) curl -o $fil http://hea-www.harvard.edu/AstroStat/Tutorial2014/$fil
  ls -l $fil # if you want confirmation of download
end
```

I had no joy with curl. Maybe their server was just getting hit too hard:

```
neptune$ curl http://astrostatistics.psu.edu/MSMA/datasets/SDSS_QSO.dat -o
outfile
  % Total    % Received % Xferd  Average Speed   Time    Time     Time
Current
                                 Dload  Upload   Total   Spent    Left  Speed
 28 8544k   28 2448k    0     0  80496      0  0:01:48  0:00:31  0:01:17 64916
curl: (18) transfer closed with 6242810 bytes remaining to read
```

Actually I had similar problems trying to load that file in the browser window.

[VK] I have downloaded all of them (in full [I think]) to hea-www.harvard.edu/AstroStat/Tutorial2014/ -- grab them from there if you are still having trouble getting them from PSU. (Open source tutorialing! Exciting!)

```
wc *.dat *.csv  # these seem to be all correct
      573    1146   8014 COMBO17_lowz.dat
      53     106    728 NGC4406_profile.dat
      59     118    840 NGC4472_profile.dat
```

```
   41     82     566 NGC4551_profile.dat
77430 1161450 8749579 SDSS_QSO.dat
   4215   21080  181545 Shapley_galaxy.dat
   5001   5001   366182 SDSS_stars.csv
12885   12885  943539 SDSS_test.csv
10091   20447  694754 SDSS_wd.csv
```

[Frank] Webcast up on working well.
Sound is good here.

[VK] How is the sound? Can you hear OK?

[PF] Really choppy here in AZ… I think I'll have to download the video files later. I'll stick around to see if it improves.

[VK] Best drop an email to CF if you continue to have streaming problems.
[PF] The mp4 stream is much better. Works fine from AZ. Thanks.

[VK] If you have questions or comments, write them out here, and Pete or I will interrupt Eric and ask on your behalf.

[Frank] So how does one sum the elements of a which are > 40?

[VK] sum(a[a>40])
just like IDL's total(a[where(a gt 40)])
and unlike IDL, won't crash if there are no elements --
> sum(a[a>max(a)])
[1] 0

[RD] I am having trouble downloading the last package, geoRR. Nevermind, is it actually called geoR? I downloaded that.

[TA] For reference the relevant IPython R magic docs are at:
http://ipython.org/ipython-doc/dev/config/extensions/rmagic.html
This refers to the latest development version of IPython (2.0.0). I need to see what's available in the stable 1.1.0.

[DWK] Is the afternoon session started? I do not see any live video, yet
[RD] Same here

[VK] Started just now.

{DWK} No audio! Now works both video and audio.
[RD] Video and audio are now working for me -- thanks!

[VK] Shaun says it will synchronize in a moment... Did it?

[RD] I keep getting errors about NaN being produced

[VK] Can you say at what stage?

[RD] It seems to be whatever I use the log10 function on y, for example:
plot(log10(x), log10(y), pch=20, col=grey(0.5),cex=0.3)

[VK] best reinitialize your x,y --

```
x <- sample(seq(0.01, 3, length.out=500))
y <- 0.5*x + 0.3^(x^2) + rnorm(500, mean=0, sd=(0.05*(1+x^2)))
xy <- cbind(x, y)
```

[RD] Thank you, that fixed it!

[RD] I am unstable to install the cobs package

[VK] there is a typo in the script. use
install.packages("cobs") ; library(cobs)

[RD] Thanks so much!

[RD] The webcast is no longer working for me

[VK] it had been stopped during the break. Should be back up now.

[RD] It’s back; thanks!

2014-jan-31

[VK] As usual, webcast viewers, if you have questions/comments, ask them here, and I will interrupt Eric to ask in your stead.

[VK] btw, there are cookies and munchies in Phillips. Please help yourselves!

[Frank] So the training sets must come from somewhere, presumably from unsupervised classification. How does one iterate to re-classify the training sets?
[VK] so the answer seems to be: you can’t. Training sets are permanent, choose them wisely!

[Frank] No audio on webcast.
[VK] sorry, it had been muted during the break. should be on now.

[PF] Is there a link where the slides are posted?

[VK] Not yet. I will put them up anon (such a useful word) and announce it on this page and also on the main Tutorial web page.
[Frank] Re: different definitions of credible ranges - aren’t highest posterior and mode-outward the same, since mode is by definition the highest posterior?
[PR] Frank, I will let Vinay address that question here after the lecture.

[VK] the mode is the highest point on the pdf, but HPD is defined by how the profile falls off. I didn't do a good job of explaining mode-out -- it starts out from the mode, but it is like Equal-tail, and can reflect back from a hard boundary.

[DWK] As Fank already requested, I would like to see the slides later.
In particular the current session, Eric seems to have a lot of good suggestions/recommendations. And I have to take off ….

[VK] btw, this talk is a dry run for the AAS Special Session Topics in AstroStatistics (Monday, Jun 2, 10am)