Studying galaxies in the nearby Universe, using *R* and *ggplot2*

Alastair Sanderson
Astrophysics & Space Research Group, University of Birmingham

www.sr.bham.ac.uk/~ajrs

Images credit: Astronomy Picture Of the Day
Overview

A taste of multivariate data visualization in an Astronomical context, demonstrating the power of R and ggplot2!

- Overview of distribution of nearby galaxies and galaxy groups
- R ideally suited to Astronomy & Astrophysics (although not yet widely used): wealth of multivariate public data (observed & simulated), free from proprietary & ethical restrictions on use
- Data from hyperLEDA galaxy database

R code snippets accompany each plot, to highlight key steps
Data structure

- Data frame with 6 columns & ~100,000 rows; each row is a different galaxy

```
gid  | Right Ascension | Declination | Velocity | Luminosity | ttype
-----|-----------------|-------------|----------|------------|------
a    |                 |             |          |            |      
a    |                 |             |          |            |      
b    |                 |             |          |            |      
b    |                 |             |          |            |      
...  |                 |             |          |            |      
```

Extract summary data frame of global group properties using 'plyr' package, e.g.

```
ddply( A, .(gid), summarize, sigma = sd(velocity) )
```
Galaxy velocity distribution

- Velocity (mostly) equivalent to *distance* (Hubble's Law)
- Rich in structure, with significant differences between South & North hemisphere (left/right panels)

```r
A$deccut <- cut_interval(A$dec, length=90)
qplot(x=vel, data=A, geom="histogram", binwidth=100, facets= ~ deccut)
```

A few nearby galaxies moving towards us
Galaxy distribution on the sky

East is left & West is right!

Milky Way region excluded

'cosmic web' of filaments

clusters & groups of galaxies

+ coord_map(projection="aitoff") + scale_x_continuous(trans="reverse")
Polar coordinate velocity plots

- Velocity over a thin slice in declination ('latitude'); shows 'peculiar velocities' of galaxies falling into groups and clusters (which contain lots of dark matter). Sometimes known as 'hockey puck' diagrams.

A$deccut <- cut_number(A$dec, n=2)
qplot(ra, vel, data=A, facets= ~ deccut) + coord_polar()

'finger of God' → galaxies falling into cluster/group
Galaxy morphological types

Elliptical galaxy

Spiral galaxy

-6 ← Hubble stage T type → +10

'EARLY' type

EO E7 SO

'SEA' type

Sa Sb Sc

SBa Sb Sc

Image credit: Niel Brandt's homepage

Edwin Hubble

A$morph <- factor(cut(A$ttype, breaks=c(-6, 0, 10), include.lowest=T, labels=c("Early", "Late")), exclude=NULL)

levels(A$morph)[3] <- "?"

useR! 2011

Alastair Sanderson
Missing data on morphology

- Histograms of galaxy luminosity, split by morphology (5000 < vel < 10,000)
- Unclassified galaxies (blue) are mostly faint
- Morphological classification is difficult, but can be achieved through 'crowdsourcing' initiatives like Galaxy Zoo (www.galaxyzoo.org)

```r
p ← qplot(LB, data=A, geom="histogram", fill=morph) + scale_x_log10()
p + scale_y_sqrt()                   p + geom_histogram(position="fill")
```
Galaxies held together by gravity; adds 'peculiar velocity' bias; close interactions and mergers between galaxies become possible, which can transform their properties.

Complete Local-volume Groups Sample (CLoGS) selected from a catalogue of groups identified by 'friends-of-friends' clustering in position & velocity space (Garcia, 1993, Astronomy & Astrophysics, 100, 47)

See http://www.sr.bham.ac.uk/~eos/CLoGS.html for more details (CLoGS project run by Ewan O'Sullivan at U. Birmingham)
Galaxies bound by large mass of dark matter ($\sim 10^{12} - 10^{15} \, \text{M}_{\odot}$) → Gaussian velocity distribution

Brightest Group Early-type (BGE; dashed line) is 'special' & usually found near group centre

---

Fritz Zwicky, the first person to infer the presence of dark matter (in 1933), from galaxy velocities.
Velocity distribution across a range of groups

- Composite of ~3500 galaxies in 82 groups; velocity scaled to zero mean & unit (robust) variance within each group
- Similar distributions for each type, but much narrower peak for *brightest group early types*, which live close to group centre

```r
+ geom_rug(data=BGE, alpha=0.3)
+ geom_density(alpha=0.33)
```
Quantile-quantile plot of velocity distribution

- Quantile-quantile plot to demonstrate Gaussian (normal) distribution of galaxy velocities within groups

Outliers: infalling or interloper galaxies

Use robust sdev estimator to suppress outlier bias

A <- ddply(A, .(lggnum), transform, svel = scale(vel, scale=mad(vel)))
A$qn <- qqnorm(A$svel, plot=FALSE)$x

qplot(x=qn, y=svel, data=A) + coord_equal() + geom_abline(intercept=0, slope=1, linetype=2)

Force equal size in x & y
Galaxy spatial distribution in groups

- Spatial density peaked & roughly circular
- BGE (‘+') usually near centre of isodensity contours
- Early-type galaxies preferentially found in highest density regions

qplot(ra, dec, data=gdf, size=LB, colour=svel, shape=morph, alpha=0.5) + scale_size(trans="log10") + geom_density2d(aes(group=1), legend=FALSE) + ...

LGG 338, aka NGC 5044 Group
T-type morphology vs. radius

- Early-type galaxies found in denser environments: 'morphology-density relation' (Dressler, 1980, Astrophys. J., 236, 351)

- Mergers & interactions transform spirals into elliptical galaxies

```
ggplot(data=A, aes(x=r, y=ttype, size=LB)) + geom_point(alpha=0.5, shape=1) + scale_size(trans="log10") + geom_smooth(legend=FALSE) + scale_x_continuous(trans="sqrt")
```

'rug' of unclassified galaxies
Dashboard of plots

- Assemble multiple panels for each galaxy group
- Panel layout set using 'grid' package
- Dashboard function applied across all groups using \texttt{d\_ply()}, with a progress bar

\begin{verbatim}
d_ply( A, .(gid), PlotPanels, .progress="text" )
\end{verbatim}

'PlotPanels' is the user's function to create the dashboard
Diversity of group properties

similar spatial distributions, but LGG 177 (left) has only 1 early-type galaxy...
Further *R*-related reading

- See the 'ggplot2' book and the paper on the use of the 'plyr' package (*J. Stat. Soft.*, vol 40, issue 1), both by Hadley Wickham:

- Leland Wilkinson's excellent book *The Grammar of Graphics* is also well worth reading
The local galaxy distribution is a highly structured multivariate dataset, ideally suited for analysing & visualising with R and ggplot2.

Roughly half of all galaxies live in groups & clusters, bound by gravity from dark matter, where interactions can change their properties.

*R is a powerful tool for tackling major unsolved problems in Astronomy & Astrophysics, especially in the era of big data...*