Separating Signal from Background

Bodhisattva Sen[1]
Department of Statistics
Columbia University

July 17, 2009

[1] Collaborative work with Matthew Walker, Mario Mateo & Michael Woodroofe

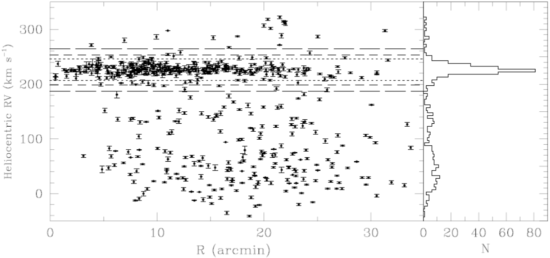
Outline

Problem

- Most astronomical data sets are polluted to some extent by *foreground/background* objects (“contaminants/noise”) that can be difficult to distinguish from objects of interest (“member/signal”)
- Contaminants may have the same apparent magnitudes, colors, and even velocities as members
- How do you *separate* out the *“signal”* ?

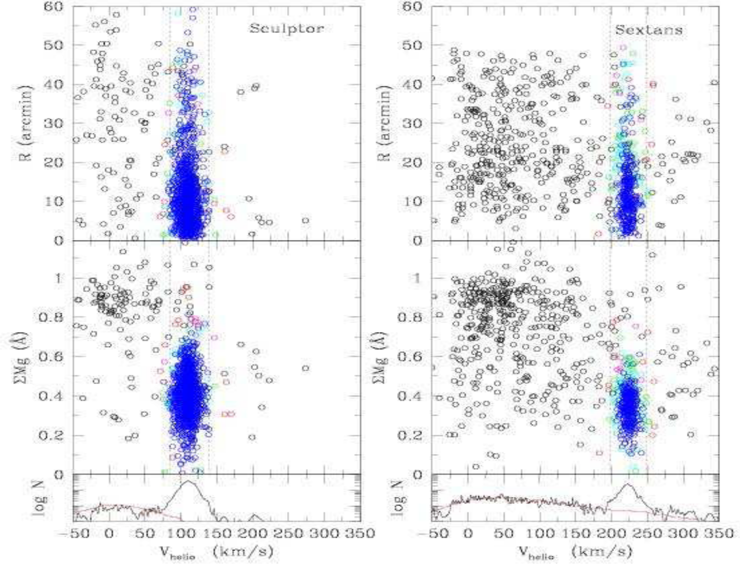
Example

- Data on stars in nearby dwarf spheroidal (dSph) galaxies
- Data: $(X_{1i}, X_{2i}, V_{3i}, \sigma_i, \Sigma Mg_i, \ldots)$
- Velocity samples suffer from *contamination* by foreground Milky Way stars



Approach

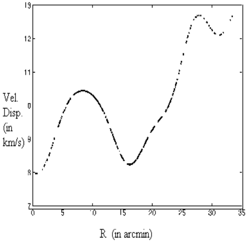
- Our method is based on the *Expectation-Maximization* (EM) algorithm
- We assign *parametric distributions* to the observables (*mixture* distribution); derived from the underlying physics in most cases
- Form the *likelihood*; can be maximized by using the EM algorithm
- The EM algorithm provides *estimates* of the unknown parameters (mean velocity, velocity dispersion, etc.)
- Also, *probability* of each star belonging to the signal population

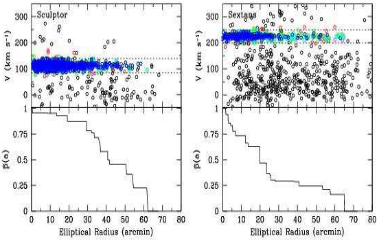

Sculptor
Sextans
R (arcmin)
ΣMg (Å)
log N
V_helio (km/s)
60
50
40
30
20
10
0
1
0.8
0.6
0.4
0.2
−50
0
50
100
150
200
250
300
350

Flexible Modeling

- Introduce *non-parametric* components
- *Velocity dispersion* was assumed constant; now can model it as a *function* of *projected radius* R

- Do not assume *exponential* density profile
- Assume that as you move from the center of the galaxy, the chance of observing a "member" star *decreases*

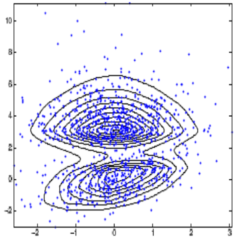


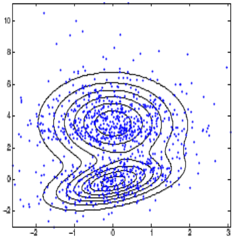


- What happens when there are *many* groups?
- Can use *Gaussian mixture models* with the EM algo
- Drawback: Each component will be elliptically symmetric

An alternative

- Mixture model $f(x) = \sum_{j=1}^{k} \pi_j f_j(x)$; $\sum_{j=1}^{k} \pi_j = 1$
- Model $f_1, f_2, \ldots, f_k$ as *log-concave* densities
- No *Tuning* parameter required; completely *non-parametric*





A further generalization

- 1-10 million data points; in arbitrary number dimensions; 100-1000 groups; highly *anisotropic* structures
- Example: identifying *substructures* in the stellar halos
- Data with dimensions of different types (apparent magnitude, angular position, radial velocity, proper motion, abundance- space) and with varying error scales
- *Clustering* procedure; e.g., *hierarchical* clustering algorithms, *break* down

A Group finder

- Sanjib Sharma and Kathryn Jonhston (2009)
- They describe a *computationally* fast, efficient group finding algorithm to identify clusters in such data sets

- Uses a *locally adaptive* distance metric
- In general, how can we handle such complex situations?

References

- Walker et al. (2009): *Astronomical Journal*
- Sen et al. (2009): *Statist. Sinica*
- Cule et al. (2009): submitted

Thank you!