Separating Signal from Background

Bodhisattva Sen
Department of Statistics
Columbia University

February 6, 2009

---

1 Collaborative work with Matthew Walker, Mario Mateo & Michael Woodroofe
1. Separating Signal Stars
   - Method
   - Extensions
   - Theory
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" stars?

We develop an algorithm for evaluating membership (estimating parameters & probability of an object belonging to the member population)
Example

- Data on stars in nearby dwarf spheroidal (dSph) galaxies
- Data: \((X_1, X_2, V_3, \sigma, \Sigma M_g, \ldots)\)
- Velocity samples suffer from *contamination* by foreground Milky Way stars
Our method is based on the *Expectation-Maximization* (EM) algorithm.

We assign *parametric distributions* to the observables; derived from the underlying physics in most cases.

The EM algorithm provides *estimates* of the unknown parameters (mean velocity, velocity dispersion, etc.).

Also, *probability* of each star belonging to the signal population.
Separating Signal and Foreground stars

Method
Extensions
Theory

Bodhisattva Sen
Separating Signal from Background
A toy example

- Suppose $N \sim \text{Poisson}(b + s)$ is the number of stars observed

- $s =$ rate for observing a *member* star

- $b =$ the *foreground* rate

- Given $N = n$, we have $W_1, \ldots, W_n \sim f_{b,s}$ where data

  \[
  \{W_i = (X_{1i}, X_{2i}, V_{3i}, \sigma_i)\}_{i=1}^n, \quad \text{and} \quad f_{b,s}(w) = \frac{bf_{b}(w)+sf_{s}(w)}{b+s}
  \]

- We assume that $f_b$ and $f_s$ are *parameterized* (modeled by the underlying physics) probability densities
For Galaxy stars

- The stellar density (number of stars per unit area) falls exponentially with radius, $R$
- The distribution of velocity given position is assumed to be normal with mean $\mu$ and variance $\sigma^2 + \sigma_i^2$

For Foreground stars

- The density is uniform over the field of view
- The distribution of velocities $V_{3i}$ is independent of position $(X_{1i}, X_{2i})$
- We adopt $V_{3i}$ from the Besançon Milky Way model (Robin et al. 2003), which specifies velocity distributions of Milky Way stars along a given line of sight
Extension: Scenario I

- Introduce a *non-parametric* component
- *Velocity dispersion* was assumed constant; now can model it as a function of *projected radius* $R$
- Needs a tuning parameter to find $\sigma(r)$
Scenario II

- 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
A further extension

- 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 on \( \mathbb{R}^p \)
- No Tuning parameter required – completely non-parametric
Model

- Suppose $N \sim \text{Poisson}(b + s)$ is the number of stars observed

- $s = \text{rate for observing a member star}$

- $b = \text{the foreground rate}$

- Given $N = n$, we have $W_1, \ldots, W_n \sim f_{b,s}$ where data
  \[
  \{W_i = (X_{1i}, X_{2i}, V_{3i}, \sigma_i, \Sigma M_{gi}, \ldots)\}_{i=1}^n, \text{ and}
  \]
  \[
  f_{b,s}(w) = \frac{bf_b(w) + sf_s(w)}{b + s}
  \]

- We assume that $f_b$ and $f_s$ are parameterized (modeled by the underlying physics) probability densities
Let \( Y_i \) be the indicator of a foreground star, i.e., \( Y_i = 1 \) if the \( i \)'th star is a foreground star, and \( Y_i = 0 \) otherwise.

Note that \( Y_i \)'s are i.i.d. Bernoulli\((\frac{b}{b+s})\). Let \( Z = (W, Y, N) \) be the complete data [where \( W = (W_1, W_2, \ldots, W_n) \) and \( Y = (Y_1, Y_2, \ldots, Y_n) \)].

The likelihood for the complete data can be written as

\[
L^C(\beta) = e^{-(b+s)} \frac{(b + s)^N}{N!} \prod_{i=1}^{N} \left\{ \frac{b f_b(W_i)}{b + s} \right\}^{Y_i} \left\{ \frac{s f_s(W_i)}{b + s} \right\}^{1-Y_i}
\]
Algorithm

- Start with some initial estimates of the parameter $\beta$ [in our simple example $\beta = (s, b, \mu, \sigma^2, \ldots)$]

E-step:
- Evaluates the *expectation of the log-likelihood* given the observed data under the current estimates of the unknown parameters
- Evaluate $Q(\beta, \hat{\beta}_n) = E_{\hat{\beta}_n}[l(\beta)|W, N]$

M-step:
- *Maximizes the expectation* $Q(\beta, \hat{\beta}_n)$ with respect to $\beta$
- Iterate until the estimates *stabilize* (which is guaranteed!)
Summary: EM algorithm with mixture models

- *Estimates* of unknown parameters
- Estimated *probability* that the \( i \)'th star is a member
- Allows *flexible* modeling (non-parametric) of data
- All we need is to form the *likelihood*!

References