Statistical Methods (summer term 2024)

Classification

(based on original lectures by Prof. Dr. N. Christlieb and Dr. Hans-G. Ludwig)

Dr Yiannis Tsapras

ZAH – Heidelberg

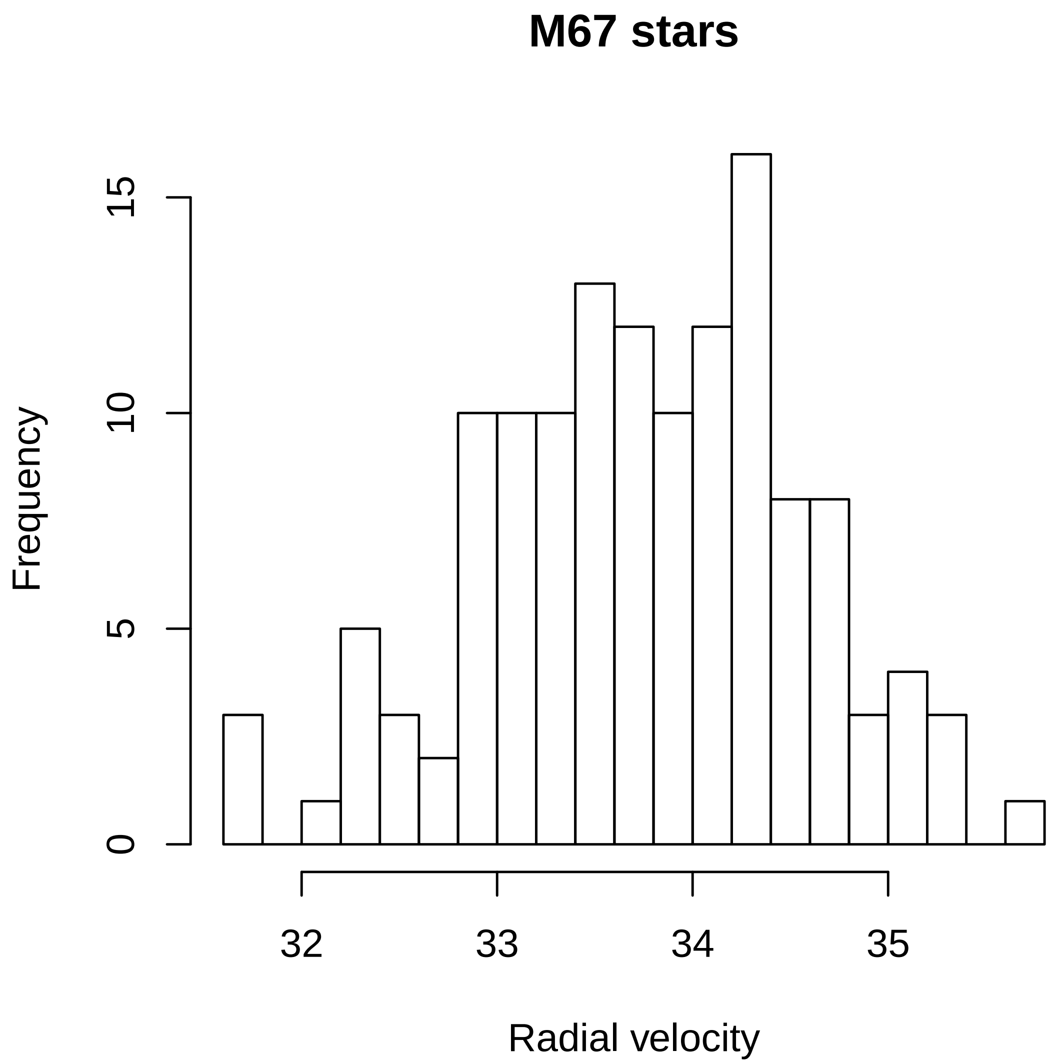
Overview

- Classification: grouping items based on their features or properties
  - A fundamental task in *machine learning*
- Example: Gaussian mixture model
  - A classic case of *unsupervised learning*
  - Relies on the EM (expectation maximization) algorithm
- Applicable to real-valued random variables (continuous data), not categorical ones

Looking at real world (or rather celestial) data

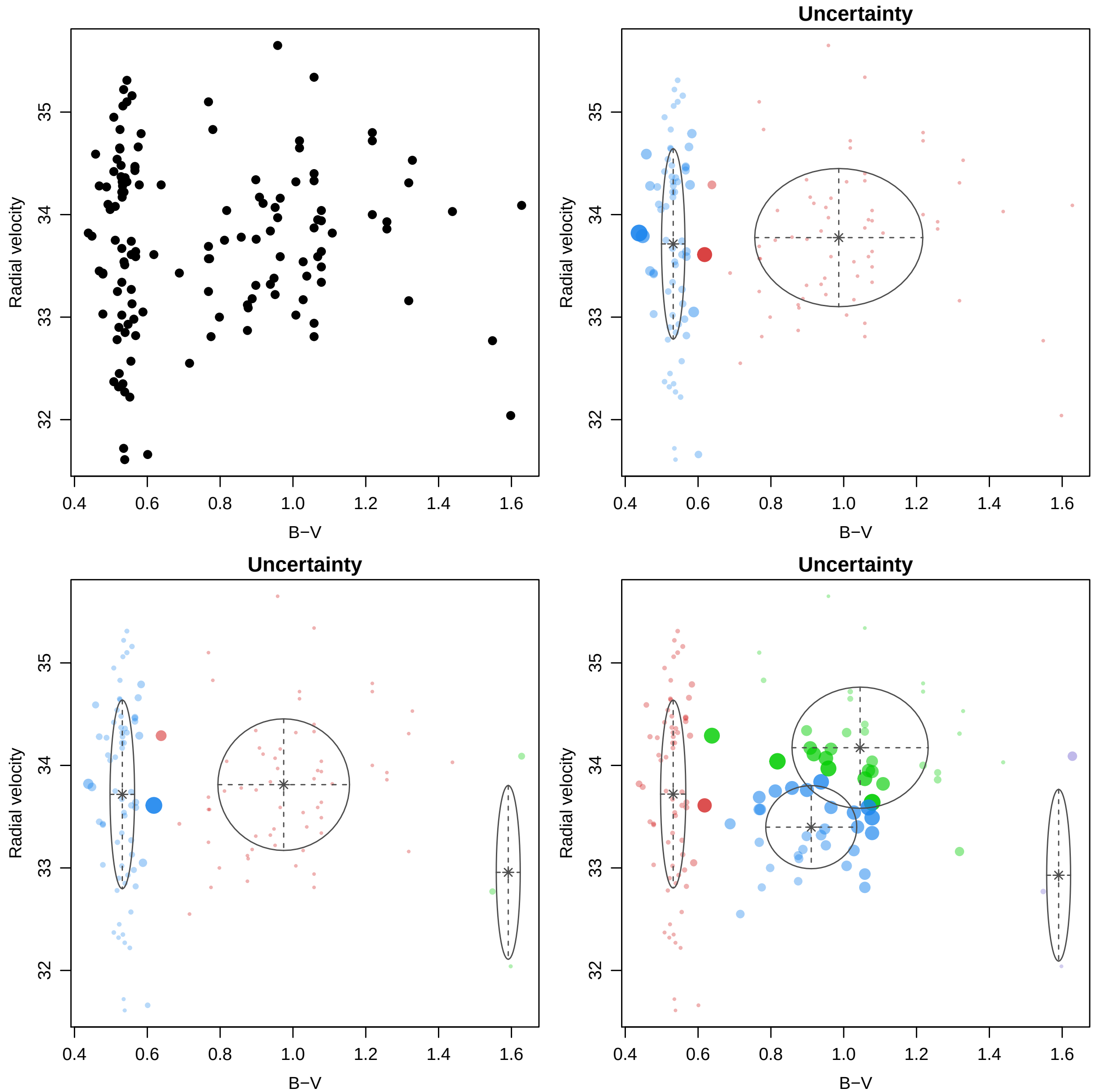
- 134 stars were observed in the open stellar cluster M67 (measurements in `stars.dat` on the web)
  - These stars are gravitationally bound and possibly formed together
  - They share nearly the same distance and *radial velocity* (RV)
- Radial Velocity (RV): The velocity component of a star along the line-of-sight, measured via spectroscopy
- RV was of interest since the gravitational redshift of stellar light was to be studied (see RV-related article on the web)
- Suspicion/Hypothesis: Dwarf and giant stars may segregate into two distinct groups based on RV

RV histogram: two groups not obviously present



How would you approach the problem to identify two groups or classes?

Here: attempt clustering analysis



Main point: stellar color B-V as further information added

Problem set-up for Gaussian mixture model

- Given: N independent data points $\boldsymbol{x}_n$ in M-dimensional space
  - Typically, M is small (e.g., 2-3 dimensions, such as RV and B-V color)
- Fitting Problem: Identify K multivariate Gaussian distributions that best describe the distribution of data points
  - Note: K (the number of Gaussian distributions) must be fixed in advance
  - The means and covariances of these Gaussian distributions are initially unknown
- Unsupervised Learning: It is not known beforehand which of the N data points belong to which of the K distributions
- Goal: Determine the N conditional probabilities $p_{nk} \equiv P(k|n)$ that point n belongs to distribution k
  - The matrix p_{nk} is known as the **responsibility matrix** (sometimes referred to as the mixing matrix)
- This responsibility matrix helps in determining how the data points are distributed among the K Gaussian distributions

Gaussian mixture model

- Things to estimate in GMM ...
  - $\vec{\mu}_k$: The mean vectors (centers) of the K multivariate Gaussians
  - $\mathbf{\Sigma}_k$: The K $M \times M$ covariance matrices of the Gaussians
  - The responsibility matrix $P(k|n)$: Probability that data point n belongs to Gaussian k

- The objective is to maximize the likelihood of the observed data:

$$\mathcal{L} = \prod_{n=1}^{N} P(\boldsymbol{x}_n)$$

Gaussian mixture model

- According to the law of total probability, the probability of each data point $P(\boldsymbol{x}_n)$ can be written as a sum over the K Gaussians:

$$P(\boldsymbol{x}_n) = \sum_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)P(k)$$

  - Here, $N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the probability density function of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$

- Typically, the EM (Expectation-Maximization) algorithm is used to maximize this likelihood

- The mixture weights p_{nk} can be computed as:

$$p_{nk} \equiv P(k|n) = \frac{N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)P(k)}{P(\boldsymbol{x}_n)}$$

- This equation provides a recipe for calculating the likelihood $\mathcal{L}$ and the mixture weights p_{nk} given the data $\boldsymbol{x}_n$

Gaussian mixture model

$$p_{nk} \equiv P(k|n) = \frac{N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)P(k)}{P(\boldsymbol{x}_n)}$$

- Problem: maximize $\mathcal{L}$ by varying the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $P(k)$
  (In a recent paper Hogg et al. worked with $N \approx 20\,000$, $M = 11$, $K = 256$)

- EM algorithm suprisingly simple and robust iterative procedure to estimate all the above parameters
  ($\rightarrow$ *Numerical Recipes* for more details of the method)

- However, there are two important considerations:
  - one must decide on the number of Gaussians K beforehand
  - as a non-linear maximization problem, the result may depend on the starting values chosen

Toying with Gaussian mixture models

- Exercise: This exercise is designed to give you hands-on experience with Gaussian mixture models and clustering analysis using real-world data
  - Step 1: Download the file `stars.dat` and the plotting routine `EMcluster.R`
  - Step 2: Load the `Mclust{mclust}` function in R
  - Step 3: Apply the `Mclust` function to the `stars.dat` dataset to explore clustering
- Explore a different dataset:
  - In R, explore available standard datasets using `library(help="datasets")`
  - Try applying `Mclust` to one of these datasets or search online for another dataset of interest
  - Select a dataset where you have some physical or contextual understanding to help interpret the results (does the grouping mean anything?)
- Interpreting Results:
  - Look at the number of clusters identified by the model and compare them with your expectations
  - Check summary statistics and plots to assess clustering quality