Mixture Models

Michael Kuhn

2017-8-26

Objectives

- Students will be able to
  - explain what is a normal mixture model
  - Students will be able to explain steps of the "Expectation-Maximization Algorithm"
  - Describe methods to select the number of components
  - apply methods from `mclust` to their own data

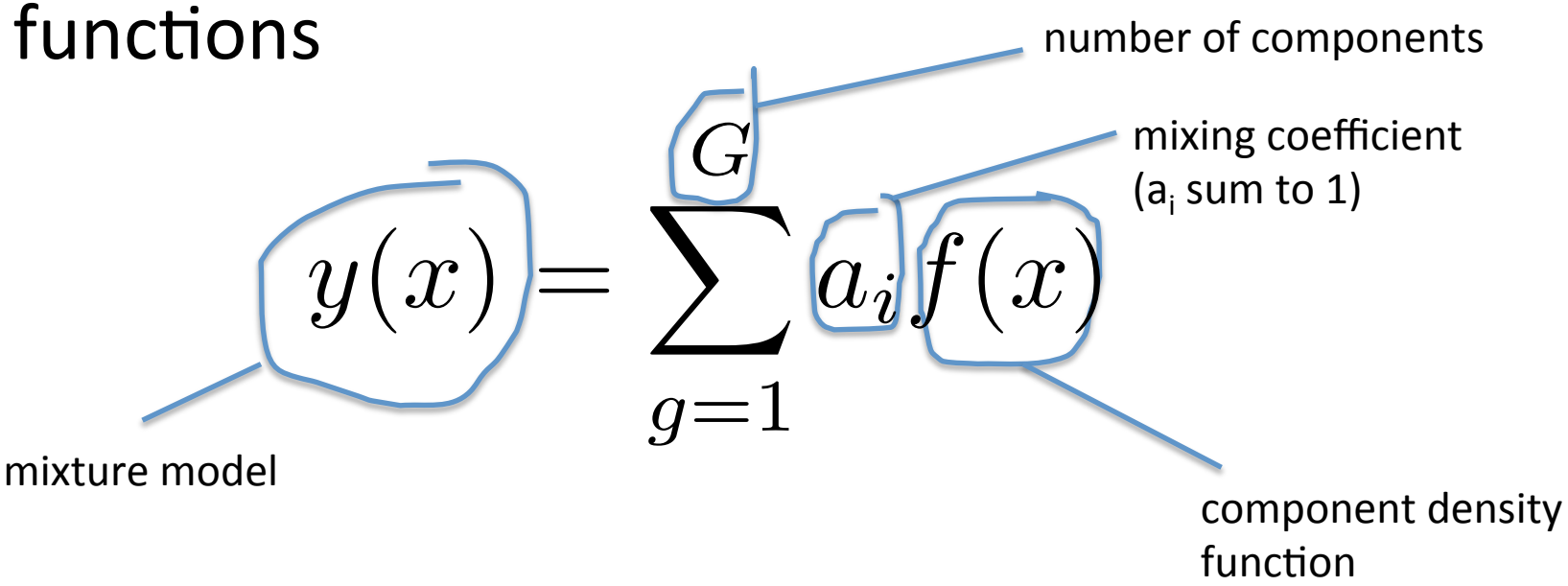
Mixture Models

- Probabilistic model for subpopulations with different probability density functions
- A probability density function generated by summing together multiple probability density functions

$$y(x) = \sum_{i=1}^{g} a_i f(x)$$

Mixture Models

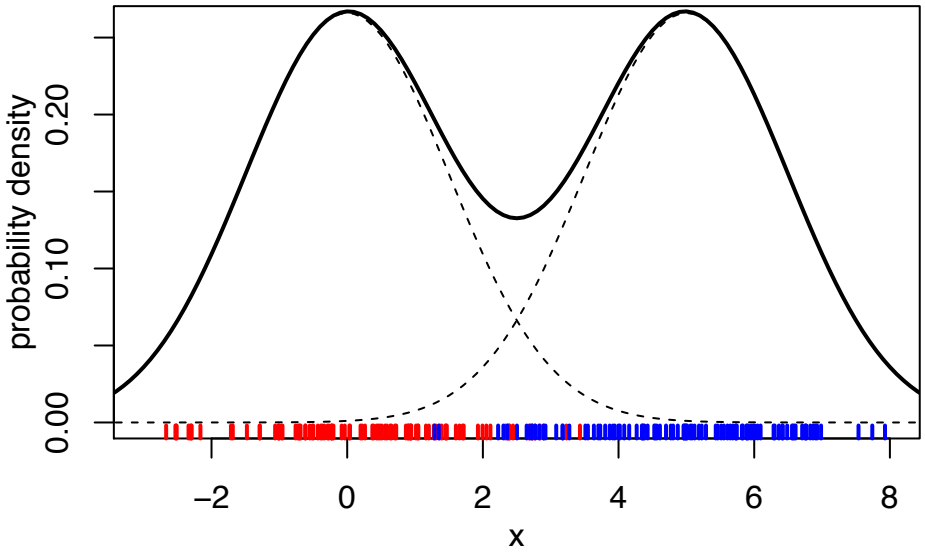
- Probabilistic model for subpopulations with different probability density functions
- A probability density function generated by summing together multiple probability density functions



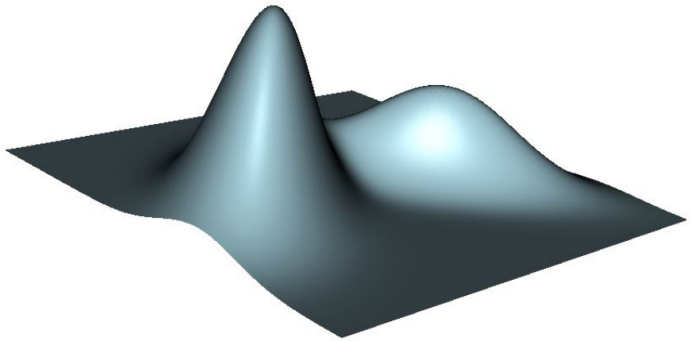
Interpretation of Mixture Models

- Arise when there is a latent variable indicating group membership

  observations (x_i, z_i)
  z_i drawn from $\{1,\ldots,G\}$
  x_i drawn from $f_z(\bullet|\theta_z)$



- Semi-parametric method of estimating density



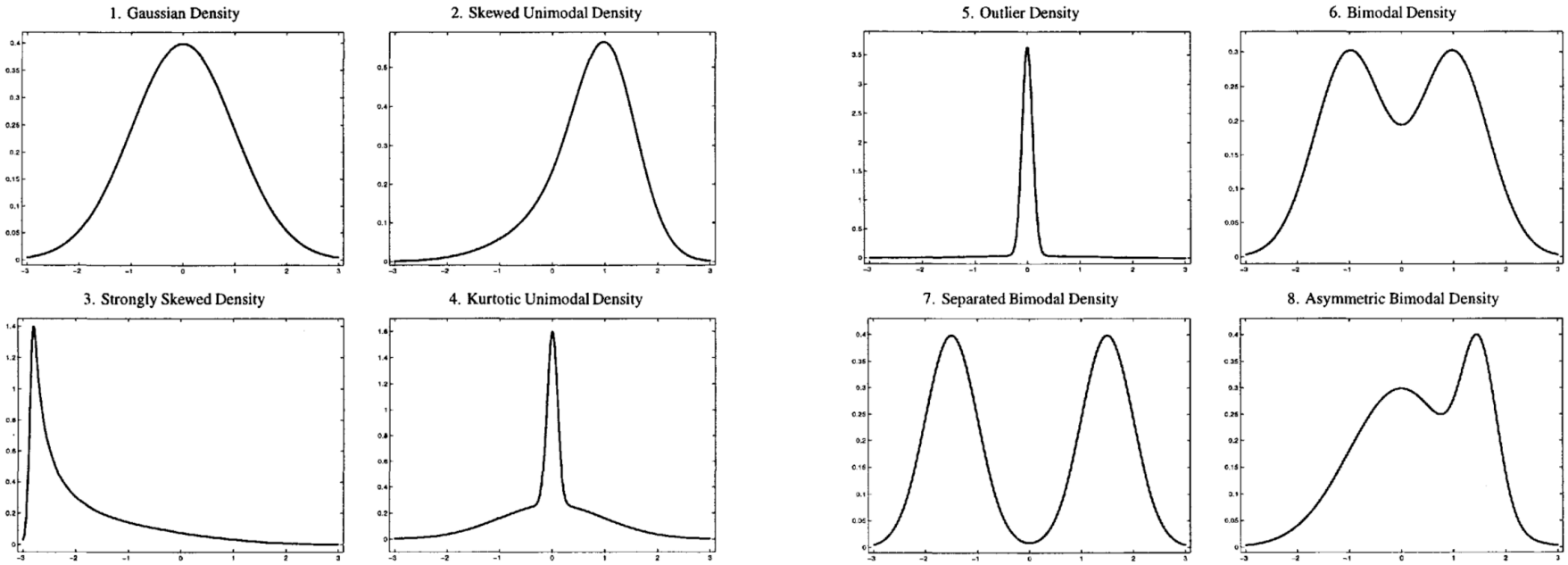
Normal (Gaussian) Mixture Models

- Normal distributions have convenient properties

univariate $$f(x|\theta) = \phi(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

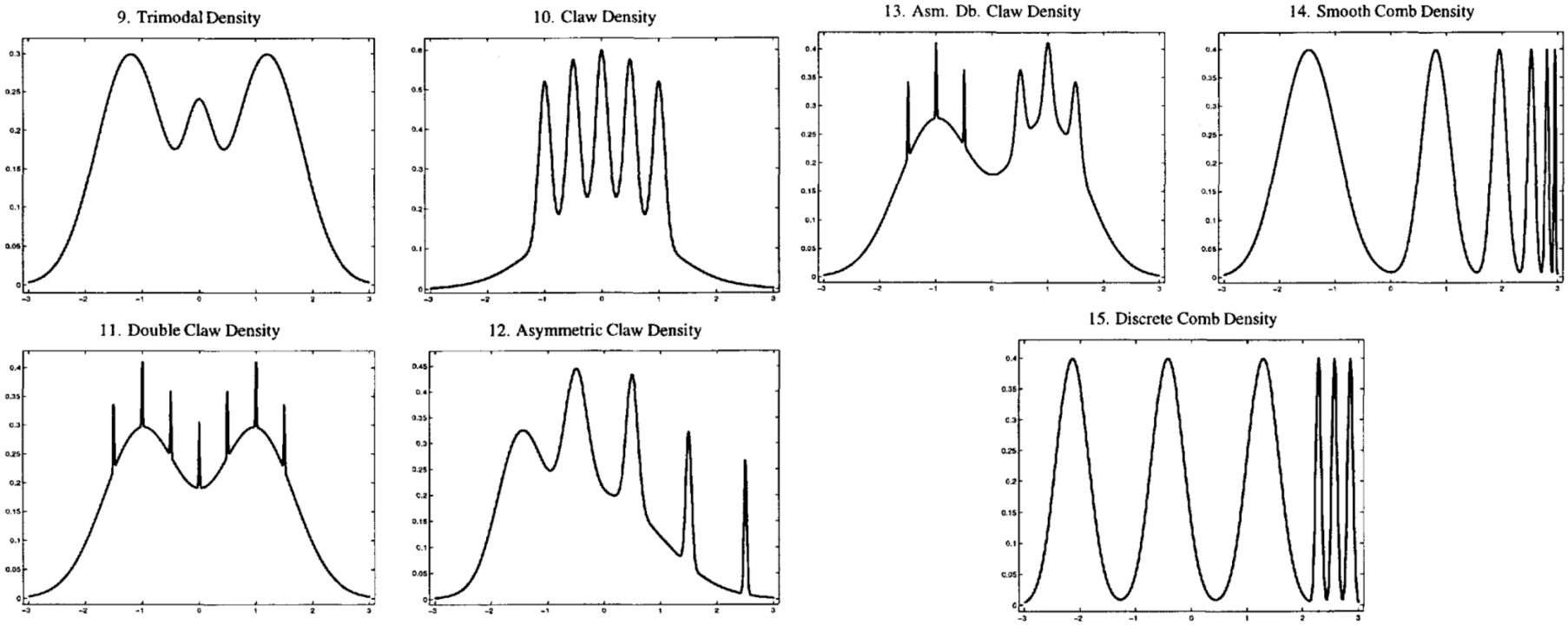
multivariate $$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$

Diversity of Normal Mixture Models



Examples from McLachlan & Peel (2004)

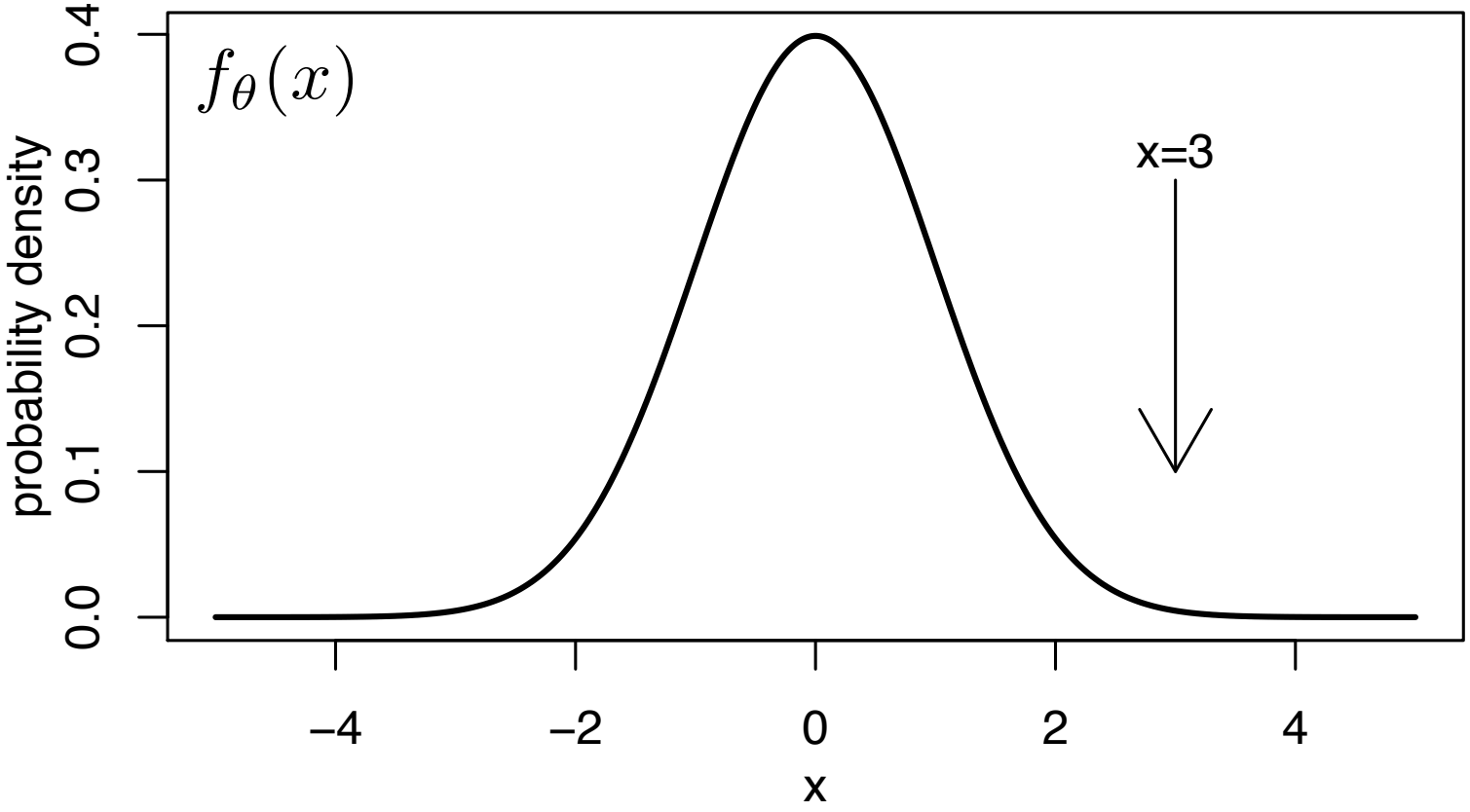
Diversity of Normal Mixture Models



Examples from McLachlan & Peel (2004)

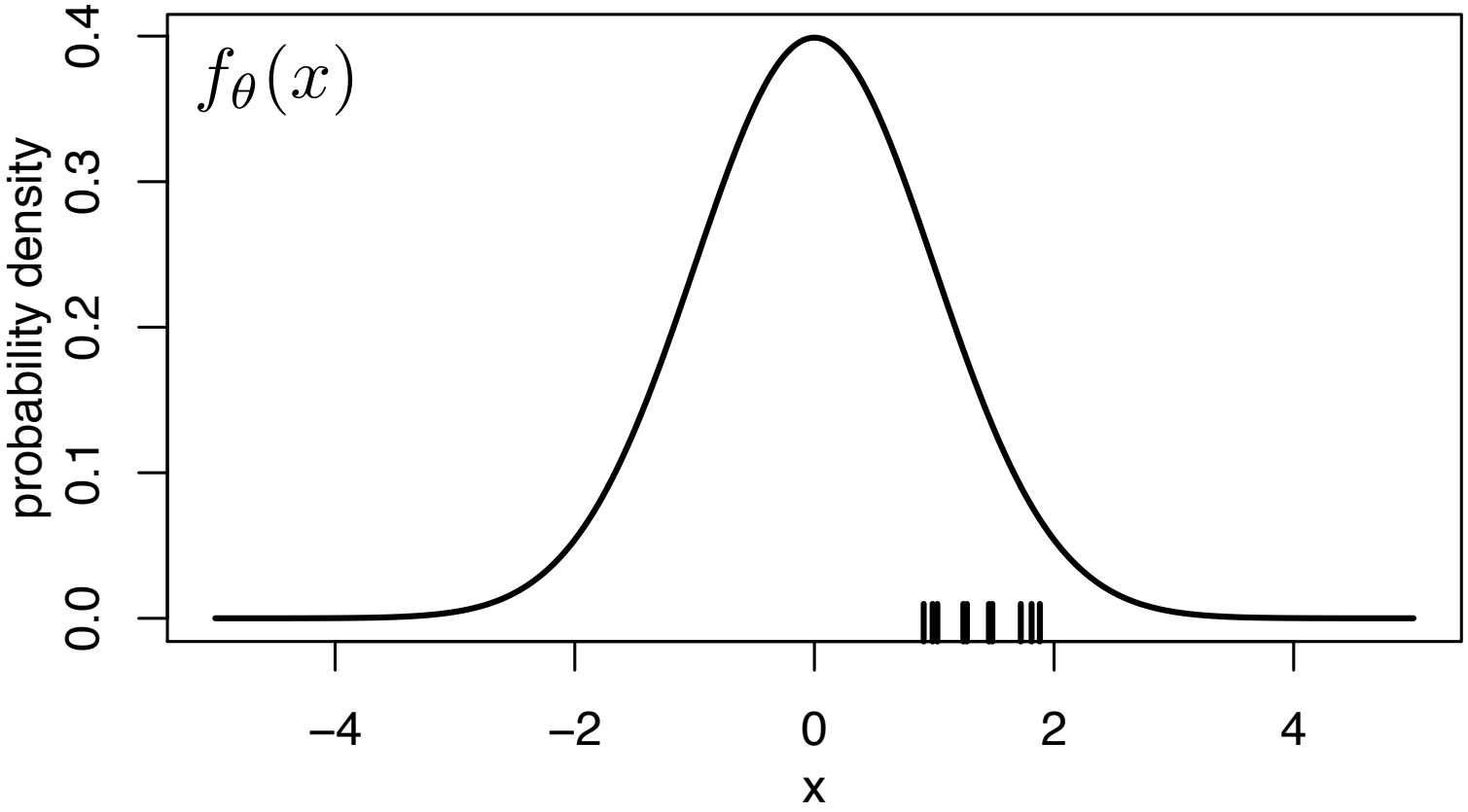
Probability

$$p(x = 3) = f_\theta(3)dx$$



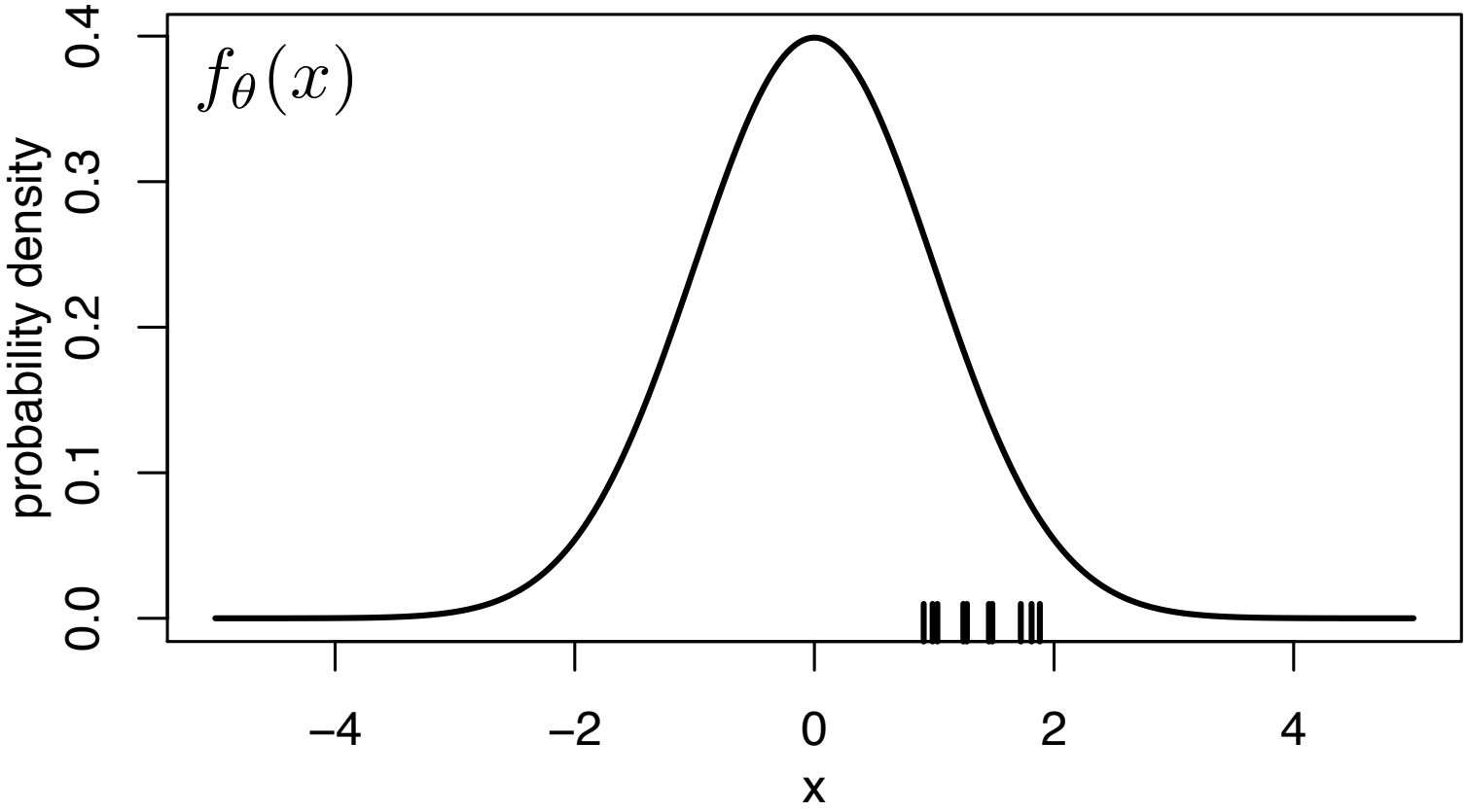
Likelihood

$$p(x_1, x_2, x_3) = f_\theta(x_1) f_\theta(x_2) f_\theta(x_3) dx_1 dx_2 dx_3$$



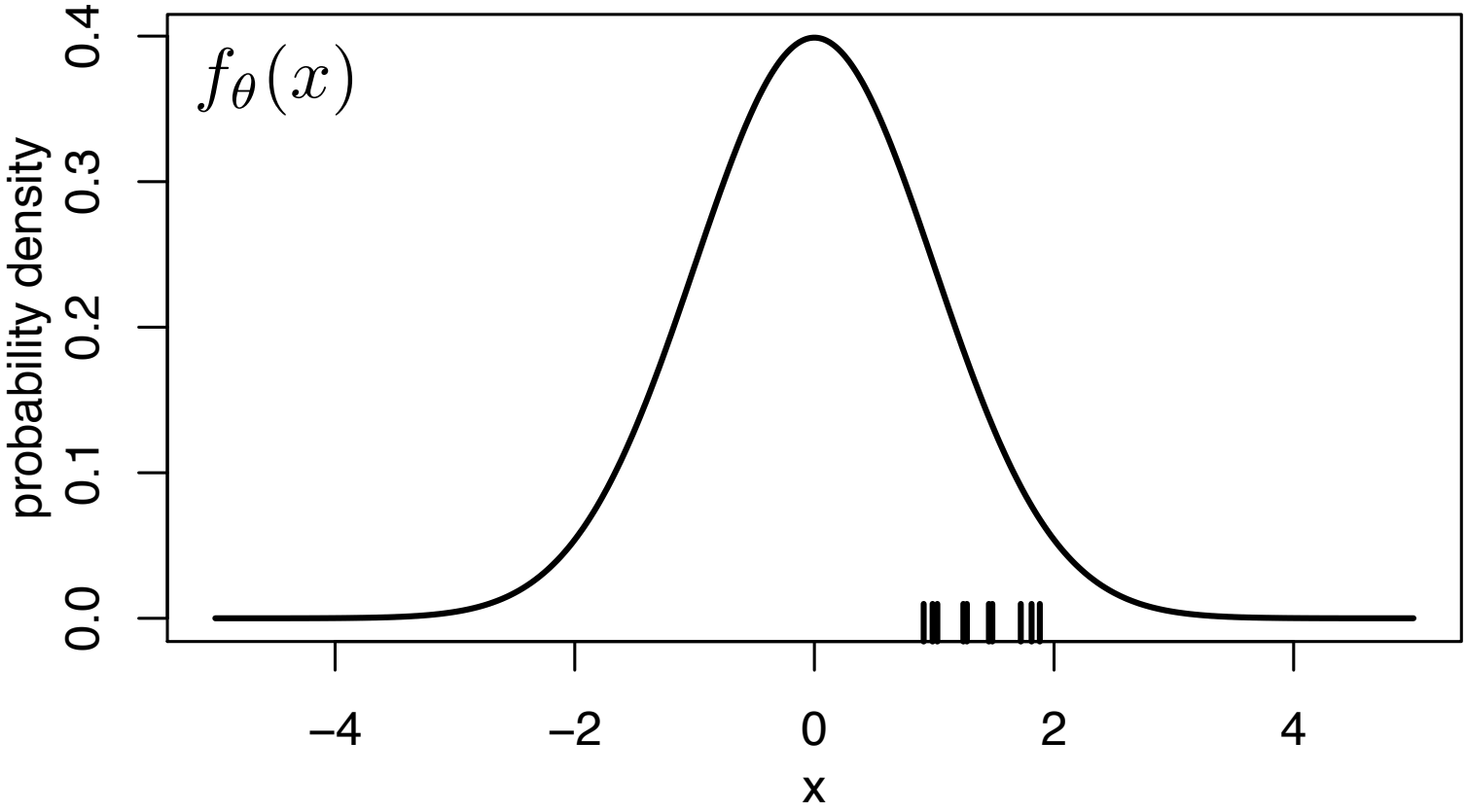
Likelihood

$$\text{likelihood}(\theta|x_1, x_2, x_3) = f_\theta(x_1) f_\theta(x_2) f_\theta(x_3)$$



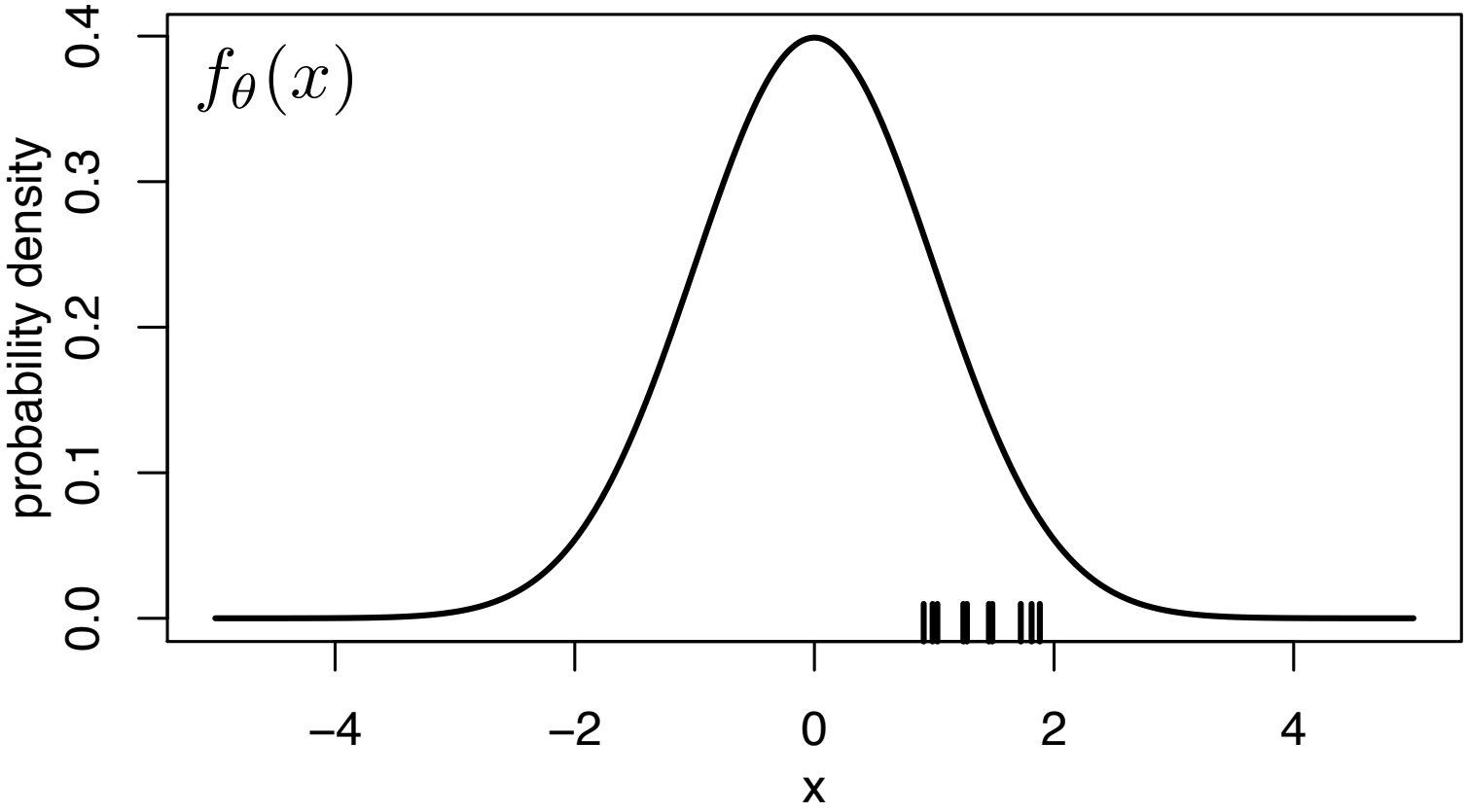
Likelihood

$$\log \text{likelihood}(\theta|x_1, x_2, x_3) = \log f_\theta(x_1) + \log f_\theta(x_2) + \log f_\theta(x_3)$$



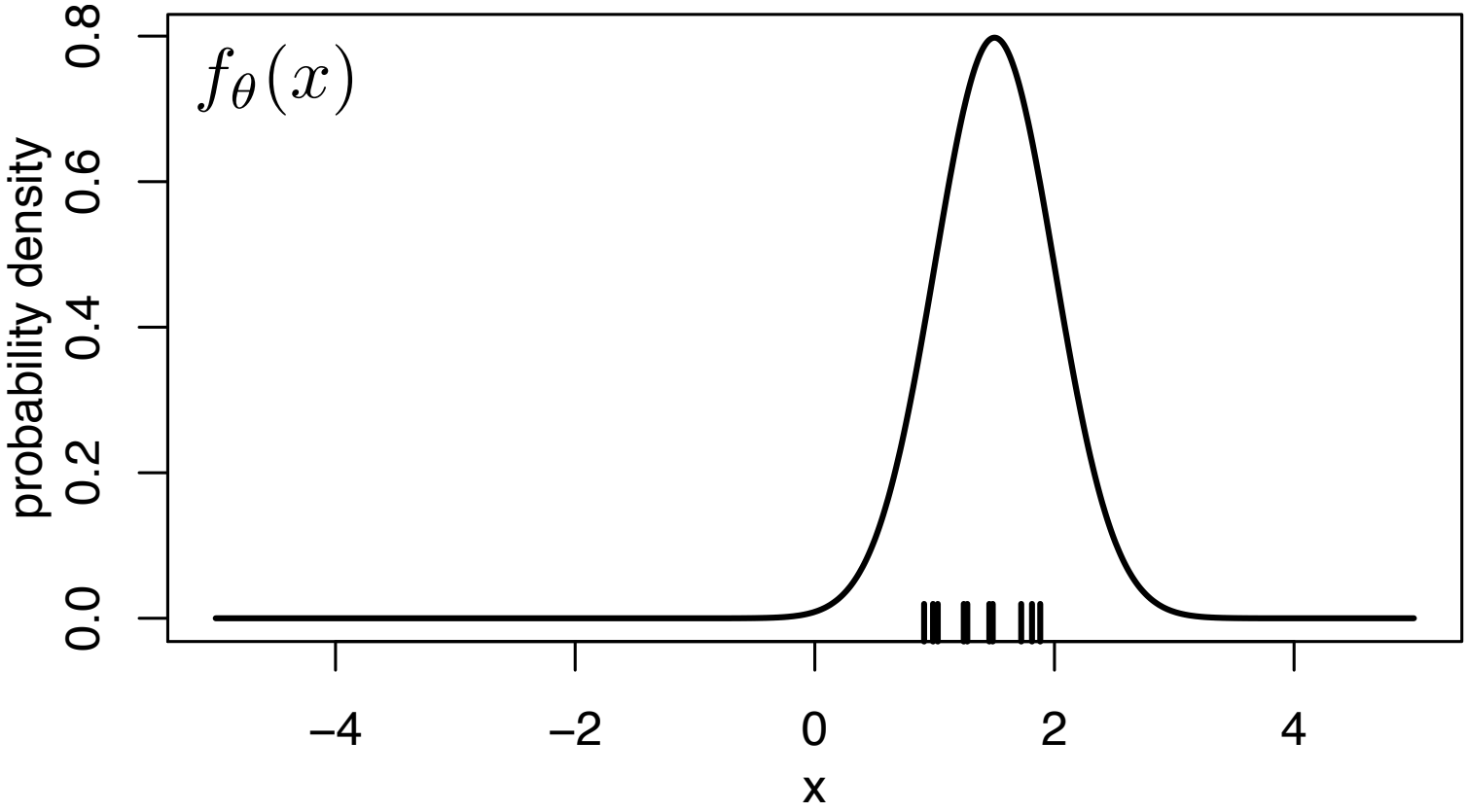
Likelihood

$$\mathcal{L}(\theta|x_1, x_2, ..., x_N) = \sum_{i=1}^{N} \log f_\theta(x_i)$$



Likelihoods

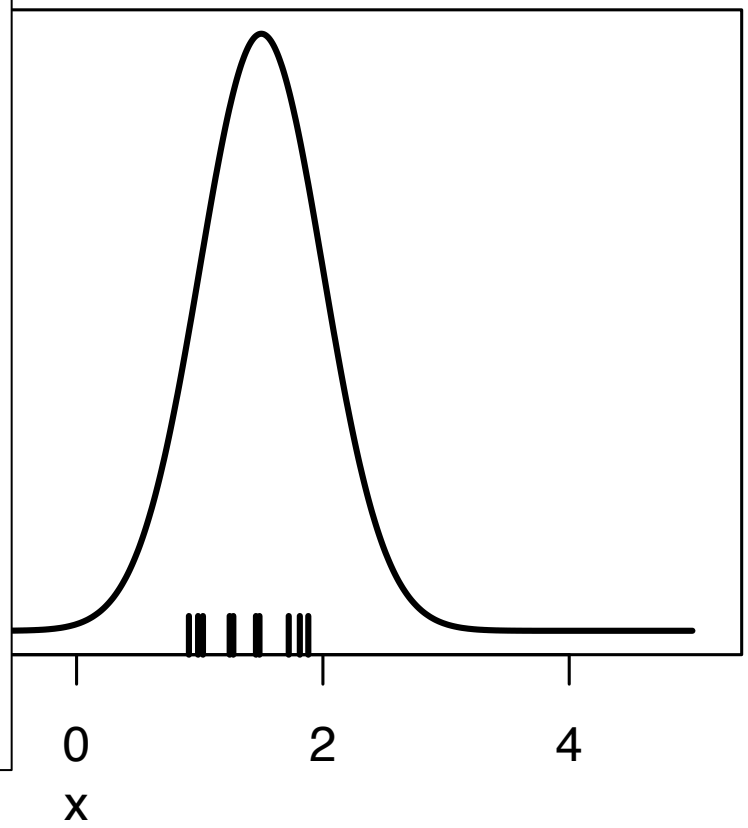
$$\mathcal{L}(\theta|x_1, x_2, ..., x_N) = \sum_{i=1}^{N} \log f_\theta(x_i)$$



Likelihoods

For a normal distribution $\phi(\mu, \sigma^2)$ the parameters are $\theta = \{\mu, \sigma^2\}$

What is the name for the maximum likelihood estimator for the center of a normal distributions?



Normal Mixture Models

- The maximum-likelihood (ML) estimates for μ and Σ can be easily estimated.

unbiased estimator of mean

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

unbiased estimator of covariance matrix

$$Q = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$$

Expectation–Maximization algorithm

- Maximum-likelihood estimation (MLE) method for a case where some variables are hidden
- In the case of mixture models, the hidden variables are the group a particular point belongs to

Expectation–Maximization algorithm

- E step: creates a function for the Expectation of the log-likelihood using the parameters from the M step
- M step: computes the parameters that maximizes the function from the E step

Expectation–Maximization algorithm

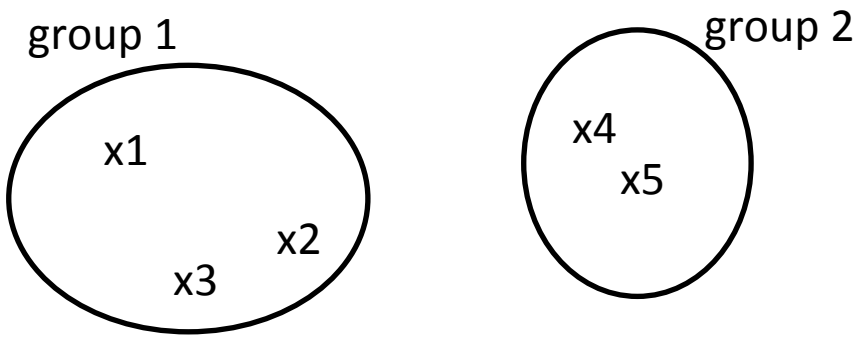
- E step: creates a function for the Expectation of the log-likelihood using the parameters from the M step
- M step: computes the parameters that maximizes the function from the E step

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\mathcal{L}(\boldsymbol{\theta};\mathbf{X},\mathbf{Z})\right]$$

Formally,

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

Expectation–Maximization algorithm



- For points $\{x_i\}$ the indicator variables variables $z_{ig} = 1$ if the i^{th} point is in the g^{th} cluster and 0 otherwise
- In general z_{ig} are not known, so we estimate the group assignments probabilistically

Likelihood for EM

- Observed Data Likelihood
  - Data are X={x_1, x_2, … x_N}

$$\mathcal{L}_{\text{obs.}}(\theta|\mathbf{X}) = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} a_i f(x_i|\theta_g) \right]$$

- Complete Data Likelihood
  - Data are X,Z = {(x_1,z_1), (x_2,z_2), … (x_N,z_N)}

$$\mathcal{L}_{\text{compl.}}(\theta|\mathbf{X},\mathbf{Z}) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log a_i f(x_i|\theta_g)$$

2 component model

$$y(x) = a_1\phi(x;\mu_1,\sigma_1) + a_2\phi(x;\mu_2,\sigma_2)$$

M step

$$a_1 = \frac{1}{n}\sum_{i=1}^{n}\Pr(1|x_i)$$

$$a_2 = \frac{1}{n}\sum_{i=1}^{n}\Pr(2|x_i)$$

$$\mu_1 = \frac{1}{na_1}\sum x\Pr(1|x_i)$$

$$\mu_2 = \frac{1}{na_2}\sum x\Pr(2|x_i)$$

$$\sigma_1 = \sqrt{\frac{1}{na_1}\sum_{i=1}^{n}(x-\mu_1)^2\Pr(1|x_i)}$$

$$\sigma_2 = \sqrt{\frac{1}{na_2}\sum_{i=1}^{n}(x-\mu_2)^2\Pr(2|x_i)}$$

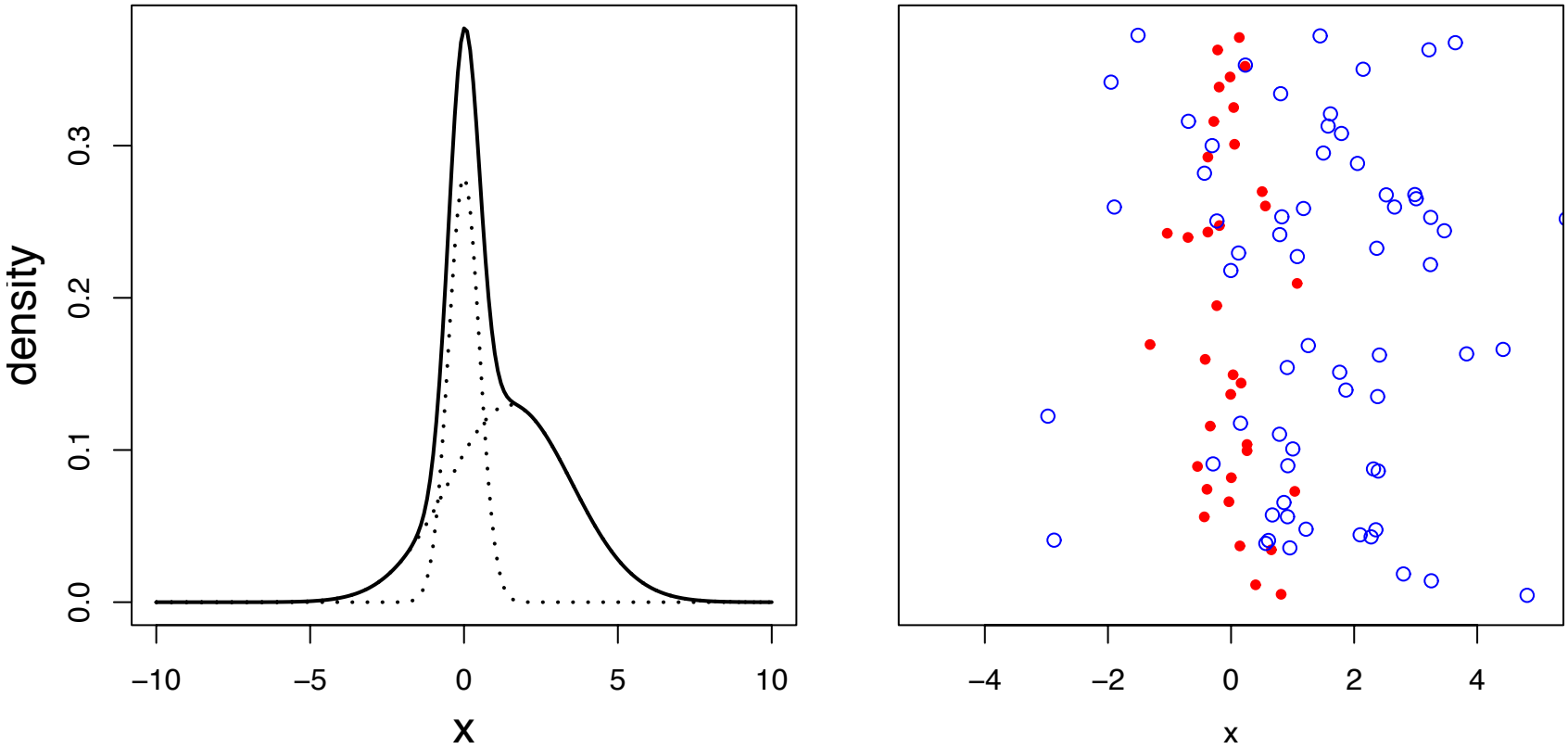
E step

$$\Pr(g|x_i) = \phi(x_i;\mu_j,\sigma_j)/y(x_i)$$

An example of the E-M algorithm

The intrinsic distribution of the simulated data

$$f(x) = 0.35\,\phi(x; \mu = 0, \sigma = 0.5) + 0.65\,\phi(x; \mu = 1.5, \sigma = 2.0)$$



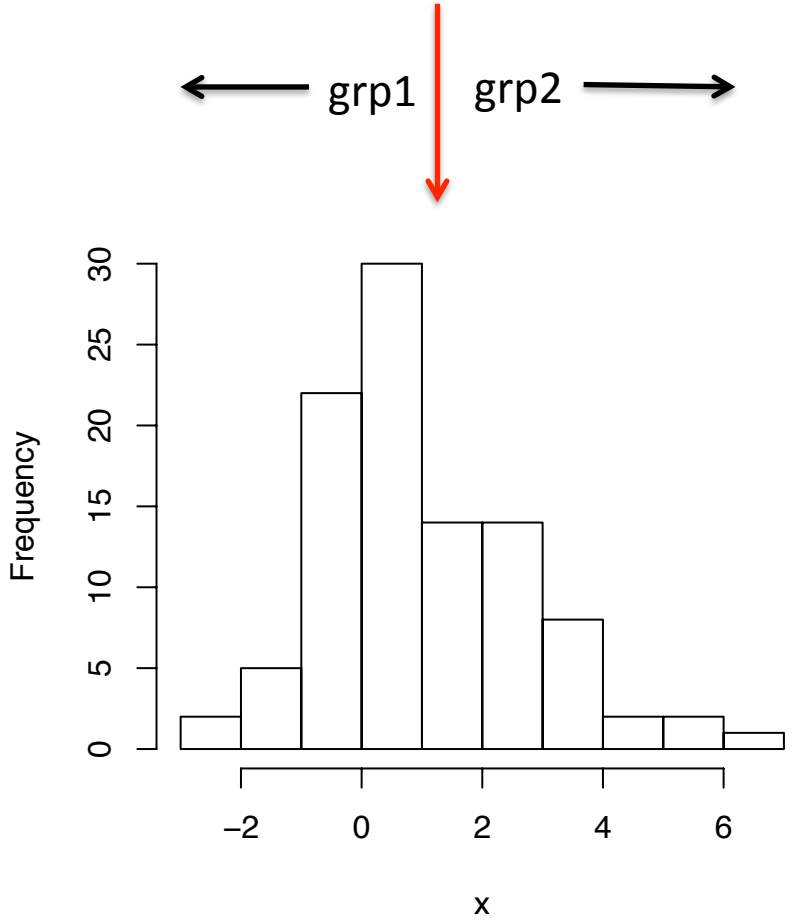
E–M algorithm: initial guess

```
n = length(x)

grp1 = which(x < 1.0)
grp2 = which(x >= 1.0)

z.indicator1 = rep(0,100)
z.indicator2 = rep(0,100)

z.indicator1[grp1] = 1.0
z.indicator2[grp2] = 1.0
```



What methods would you as an initial guess for clustering in a multivariate dataset?

M step

$$a_1 = \frac{1}{n}\sum_{i=1}^{n}\Pr(1|x_i)$$ 0.59

$$a_2 = \frac{1}{n}\sum_{i=1}^{n}\Pr(2|x_i)$$ 0.41

$$\mu_1 = \frac{1}{na_1}\sum x\Pr(1|x_i)$$ -0.1

$$\mu_2 = \frac{1}{na_2}\sum x\Pr(2|x_i)$$ 2.6

$$\sigma_1 = \sqrt{\frac{1}{na_1}\sum_{i=1}^{n}(x-\mu_1)^2\Pr(1|x_i)}$$ 0.8

$$\sigma_2 = \sqrt{\frac{1}{na_2}\sum_{i=1}^{n}(x-\mu_2)^2\Pr(2|x_i)}$$ 1.3

Number of Components

- Use methods for model selection
- Penalized likelihoods

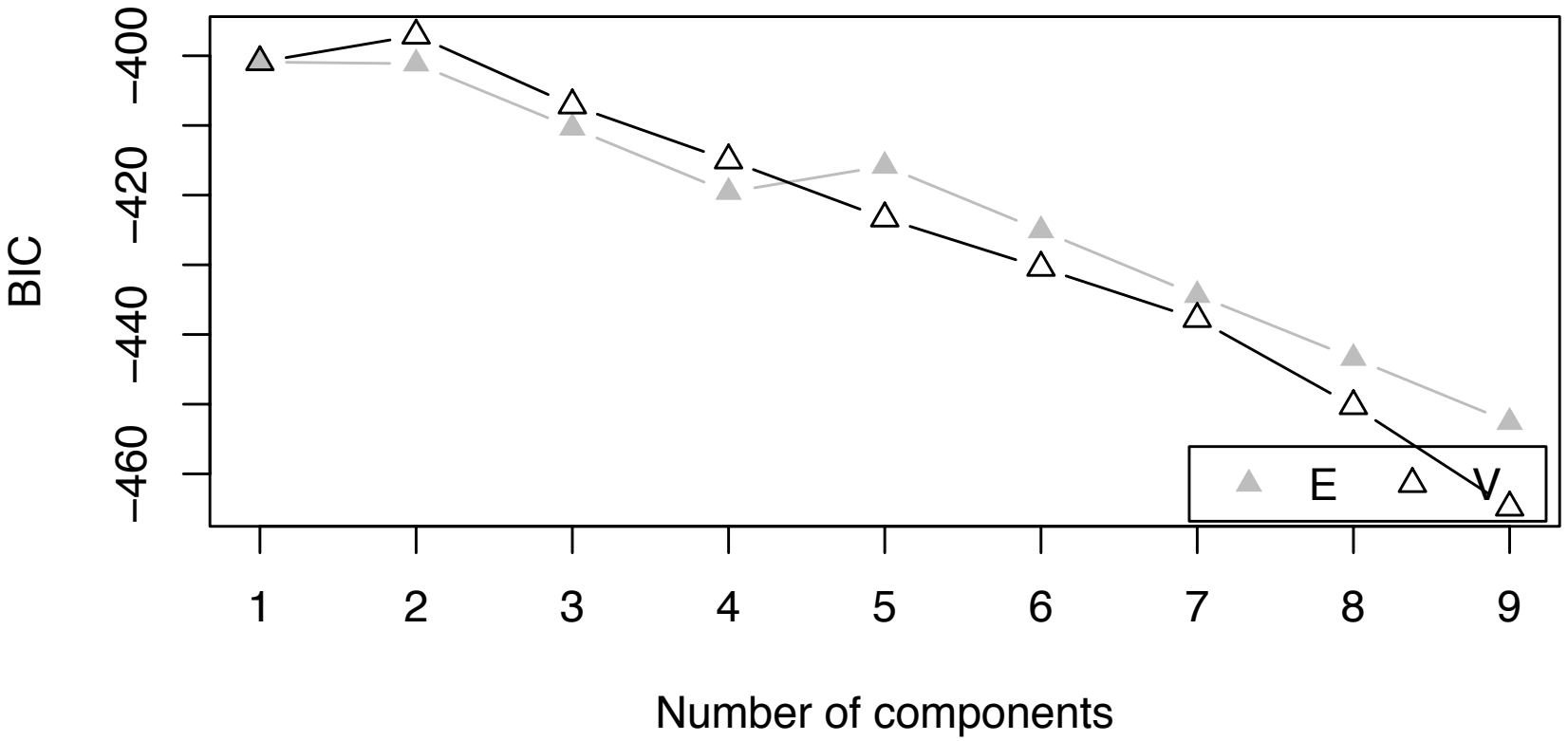
Log-likelihood: $$\mathcal{L}(\theta|\mathbf{X}) = \sum_{i=1}^{n} \ln y(x_i;\theta)$$

$$BIC = -2\mathcal{L} + k\ln(n)$$ (stricter)

$$AIC = -2\mathcal{L} + 2k$$ (less strict)

Individual values of the BIC and AIC by themselves are meaningless. They are only useful for comparison between models.

BIC



mclust

- EM algorithm for fitting models
- BIC for model selection
- Optional, restrictive or unrestrictive priors

Open R and the notes from the school webpage.

Bayesian Methods

- MCMC methods useful for parameter estimation and model selection
- Bayesian methods are particularly useful for evaluating the effect of uncertainties on model parameters

Problems

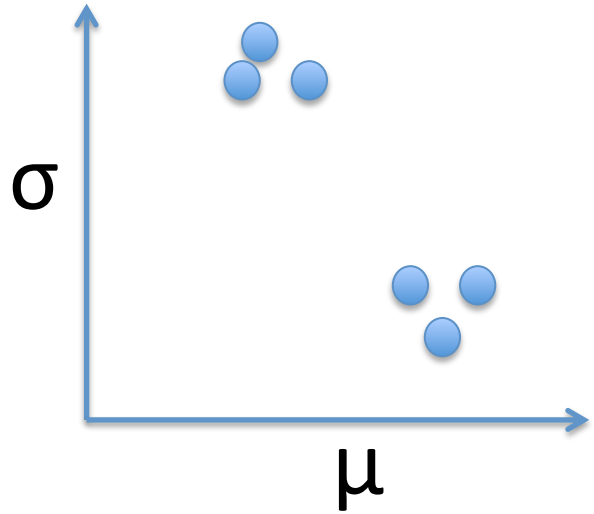
- Interpretation of results
  - Multimodality of Density
  - Nonintuitive group membership
- Fitting
  - Multimodality of likelihood
  - Infinite likelihood
  - Irregularity
  - Nonidentifability of components
  - Slow convergence

Singularities

- The likelihood of mixture models is unbound at the edge of parameter space
- A component can shrink on a given point, the mean becomes the point, and the variance goes to zero.
- Mixture model codes must implement strategies to avoid this effect

Nonidentifiability

- Permutation of labels
  - A challenge for implementations of MCMC
  - Point process representation



- Components with equal means and covariance matrices
  - Interpretation of the likelihood ratio statistic is not straightforward

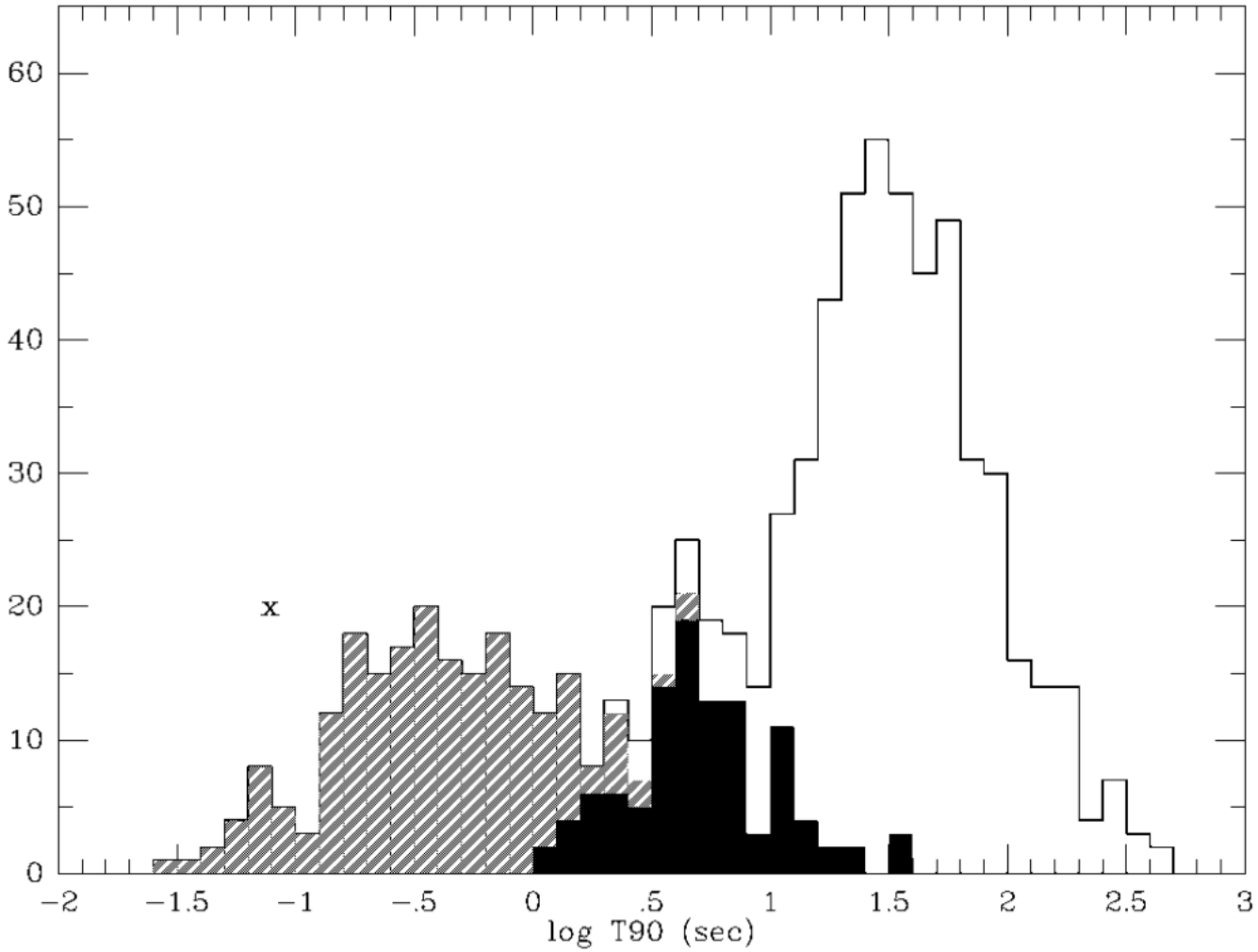
Relationship to Other Cluster Analysis Algorithms

- Normal models are a soft classifier
  - Conditional probabilities estimated
  - Classification based on estimated probabilties
- Hard classification methods
  - K-means
  - DBSCAN

Motivations for Use of Mixture Models in Astronomy

Investigating Multmodality

GRB durations from BATSE



Mukherjee et al. (1998)

log GRB Duration (T90) [sec]

Investigating Multimodality

BATSE GRBs

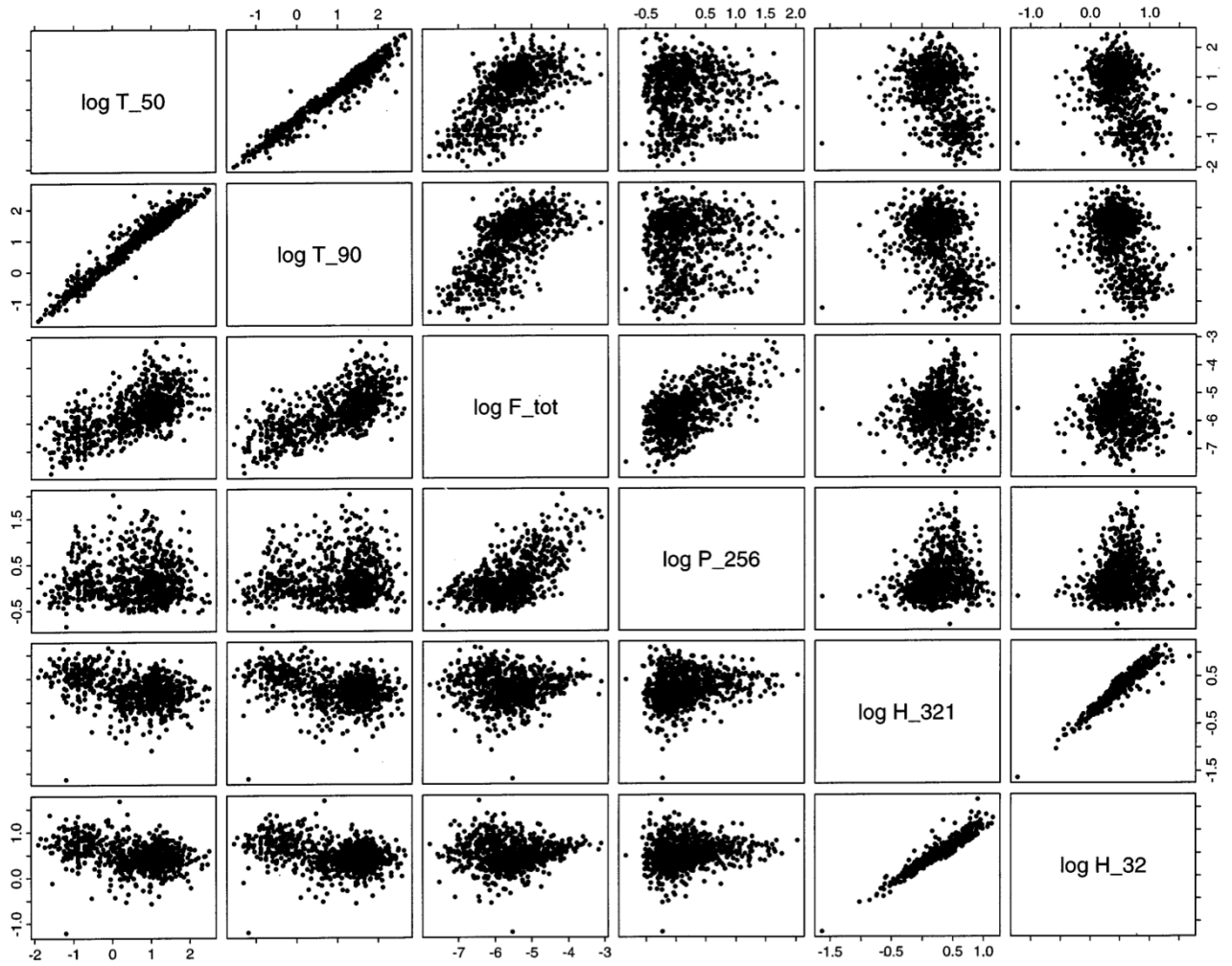
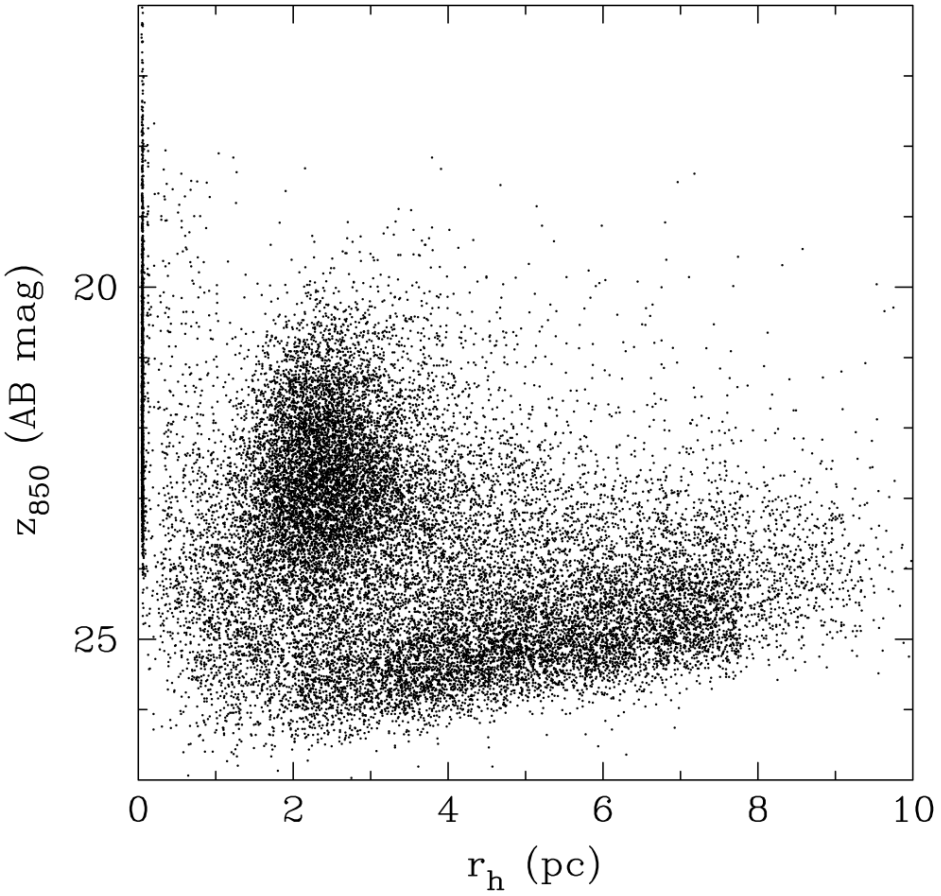


FIG. 1.—Mosaic of scatter plots of six bulk properties for the 797 GRBs from the BATSE 3b catalog used in this study

Mukherjee et al. (1998)

Removal of Contaminants

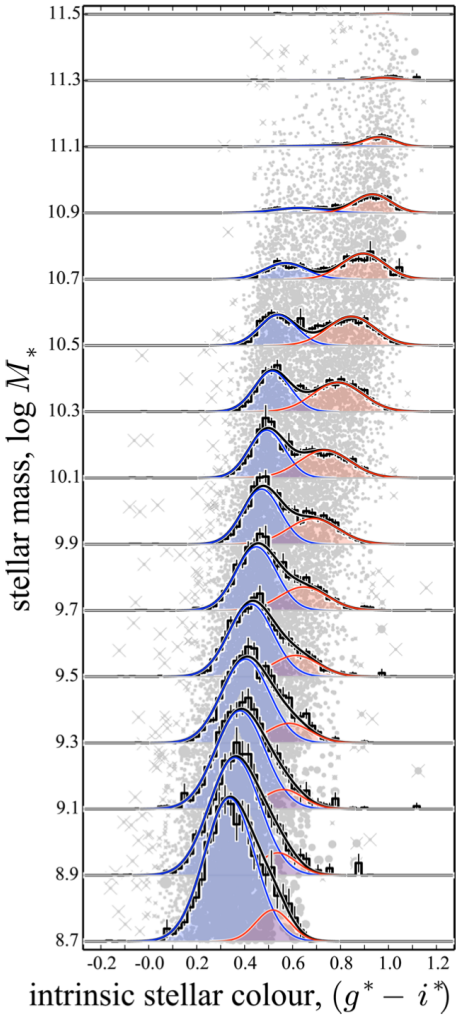
- HST observations of galaxies
  - GCs
  - Background galaxies
  - Field stars
- Mixture model approach to separate GCs and contaminants
- Distribution of galaxies obtained from an off-target field



Jordán et al. (2009)

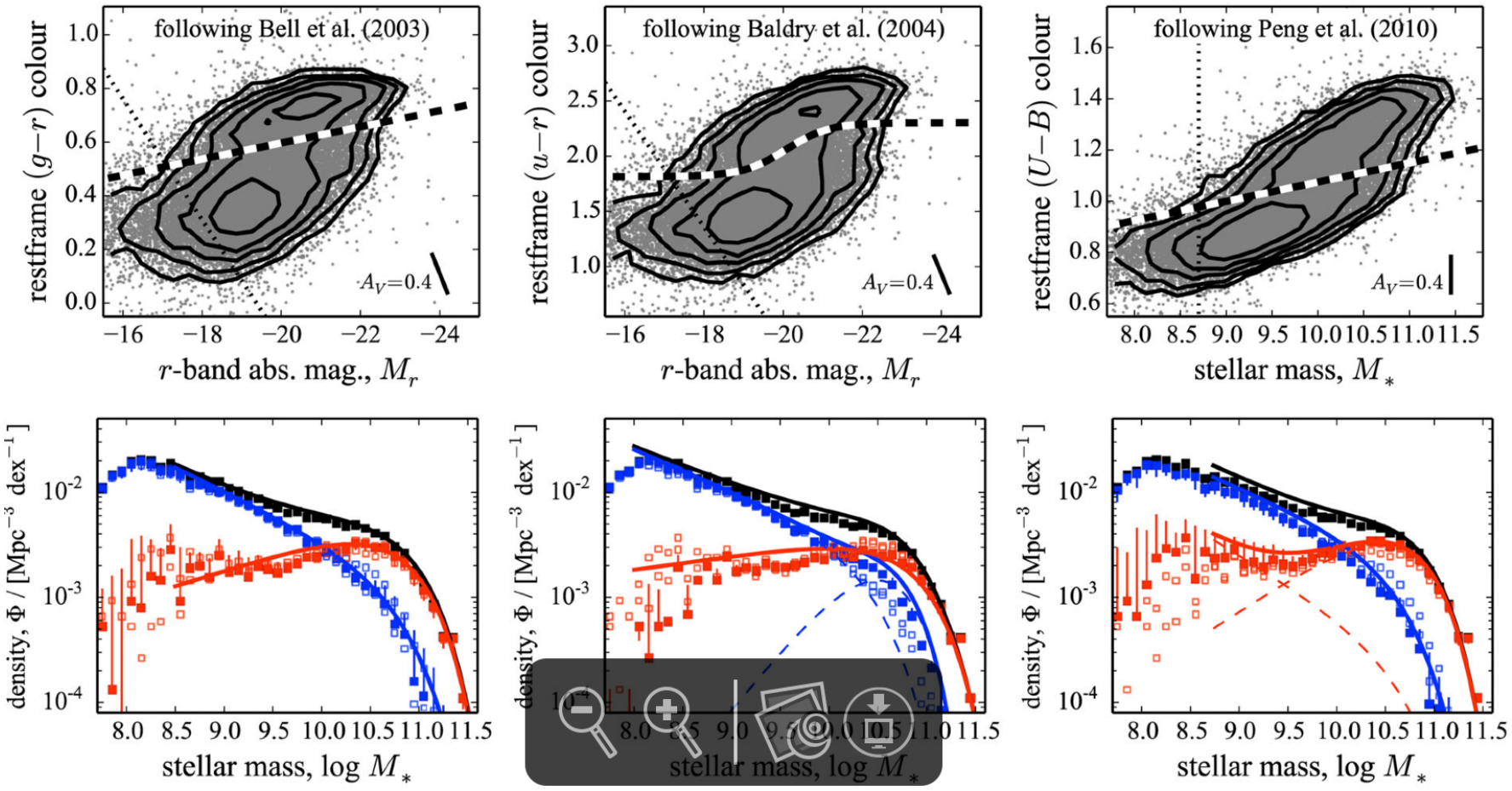
Red and Blue Galaxies

- Well-known bimodality to galaxy color distributions
- Color distributions overlap
  - Simple color cuts lead to bias when examining the galaxy mass function for "red" and "blue" galaxies



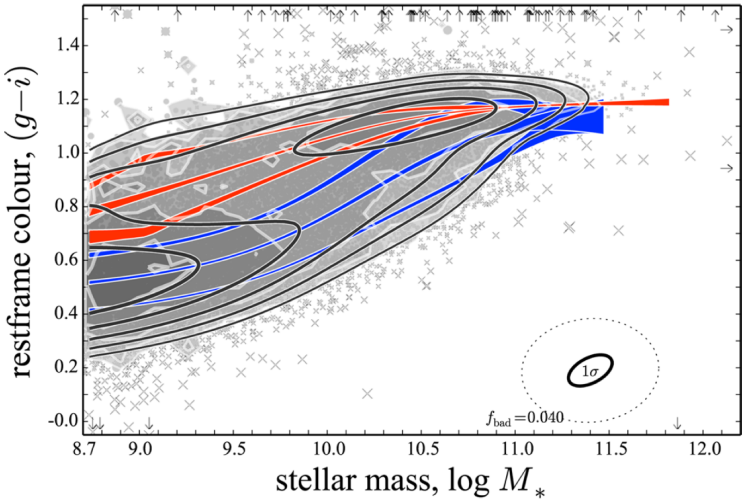
Taylor et al. (2015)

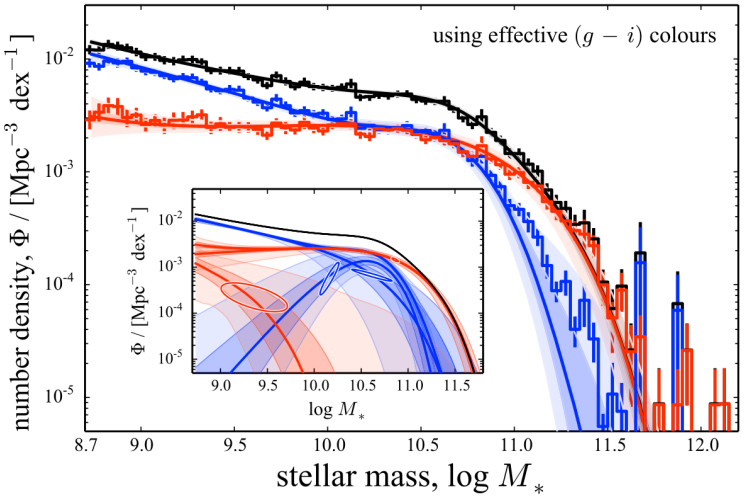
Red and Blue Galaxies



Red and Blue Galaxies

- >23,000 galaxies from GAMA
- 40 parameter model for “red” and “blue” galaxies
  - astrophysically motivated
  - MCMC
- Expected shape to mass functions
- Mixture models reveal a longer distribution of “red” galaxies that merge with “blue” population at low masses

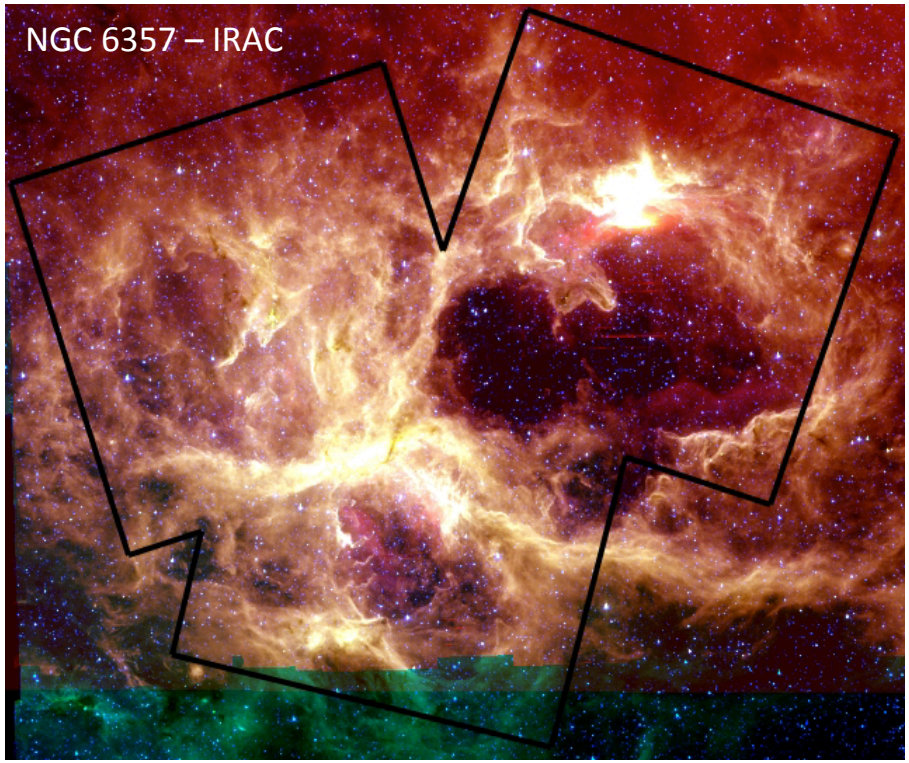


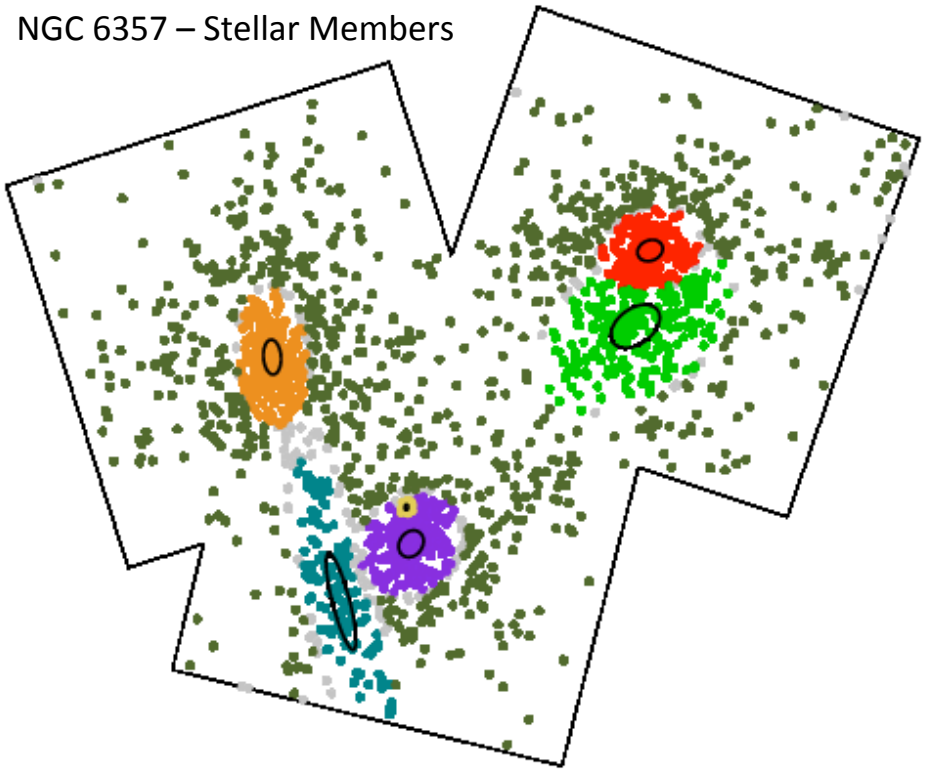


Taylor et al. (2015)

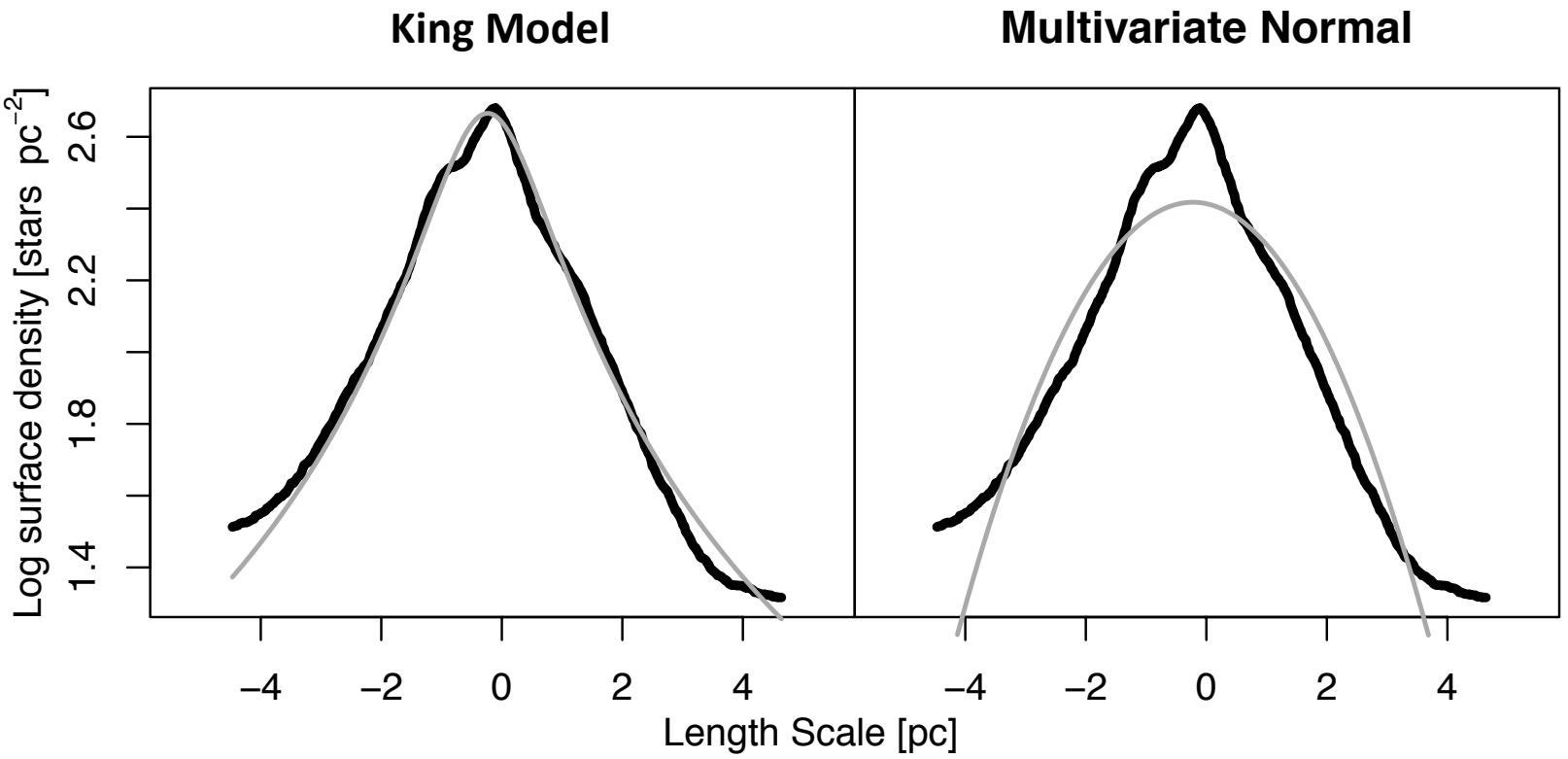
Star Cluster Modeling

- In star-forming regions young stars are often distributed in a number of groups





Star Cluster Modeling

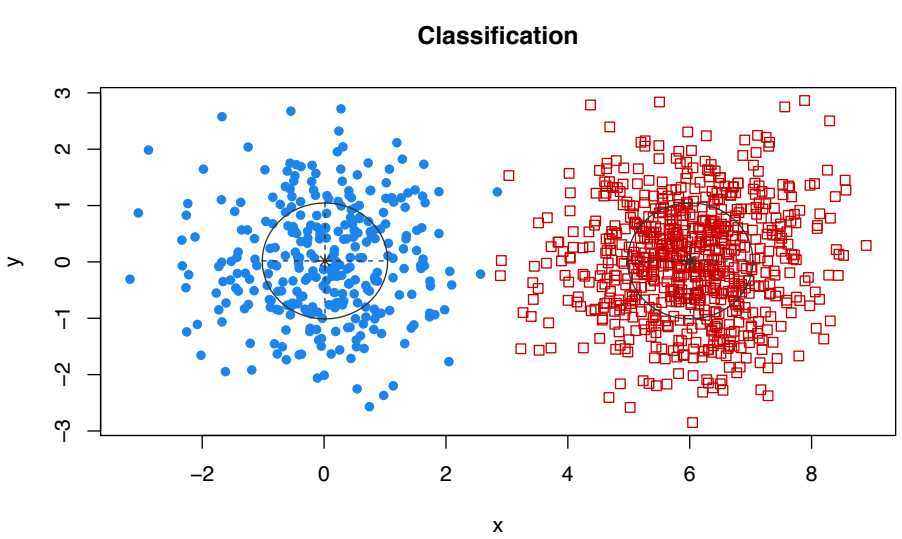


Adaptively smoothed data are shown in black

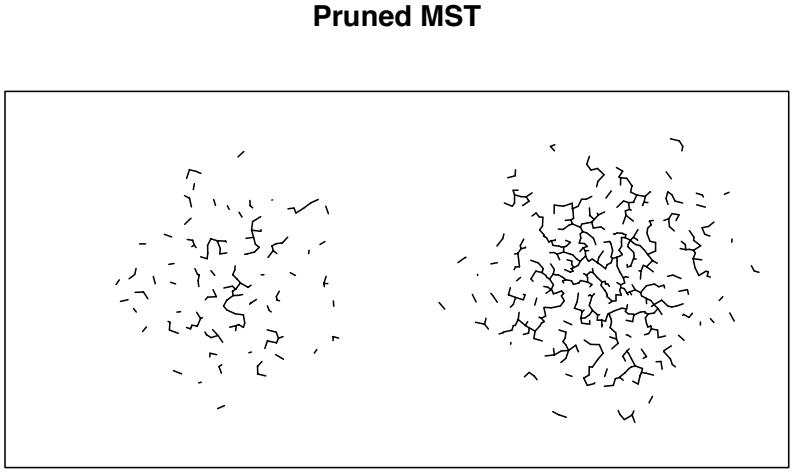
Models are shown in gray

Comparison of Methods

Mixture Model



"Minimum spanning tree" method

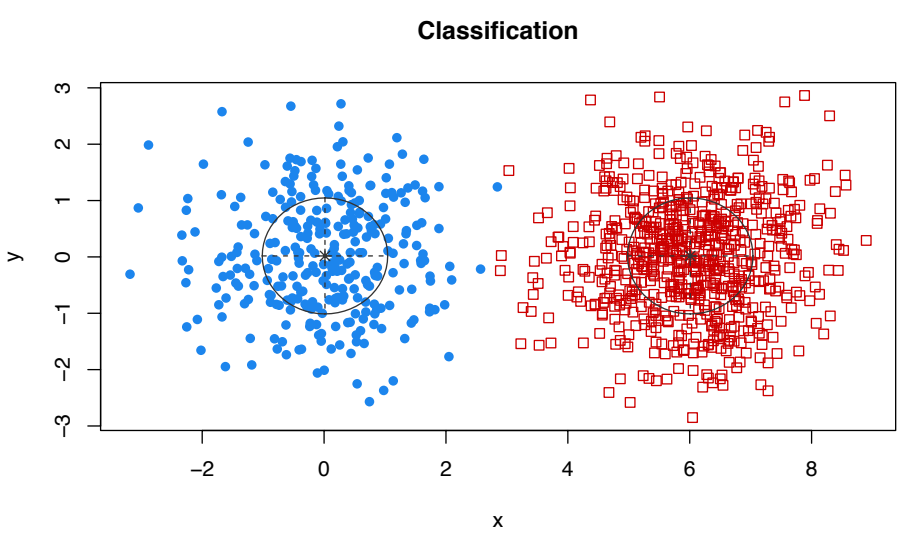


Two simulated spherical clusters

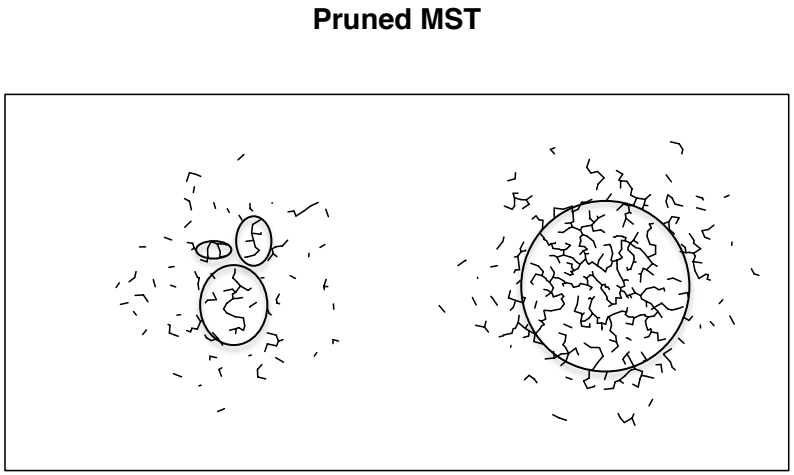
1) 300 points
2) 700 points

Comparison of Methods

Mixture Model



"Minimum spanning tree" method

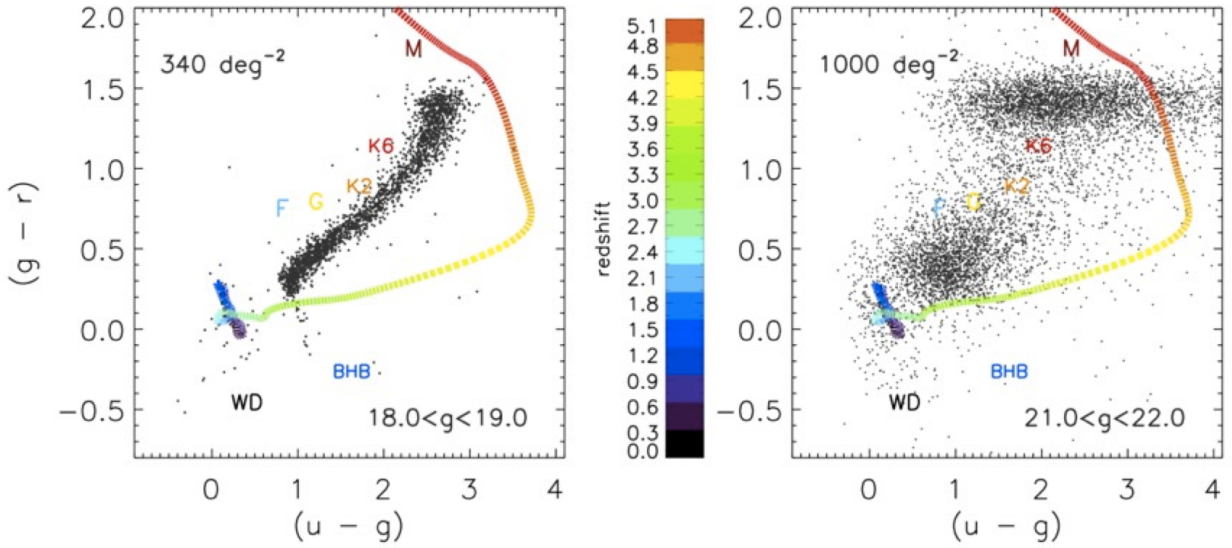


Two simulated spherical clusters

1) 300 points
2) 700 points

Extreme Deconvolution

- Technique described by Bovy et al. (2011)
- Deals with heteroscedastic errors
- Makes use of mixture models to approximate arbitrary density functions



Examples in R