Label Switching in Mixture Models

Ricky Su
University of Rochester, Data Science

May 2, 2017

Overview

Label Switching

Example

Approaches

Results

What is Label Switching?

- Arises when taking a Bayesian approach to parameter estimation, and clustering using mixture models
- Describes the invariance of the likelihood under relabelling of the mixture components
- Leads to highly symmetric and multimodal parameter posterior distributions
- Joint posterior summary by marginal posterior distributions are therefore often inaccurate

Notation

- Independent observations $\boldsymbol{x} = x_1, ..., x_n$
- k finite and known components
- Standard finite mixture model:

$$p(x|\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_1 f(x; \phi_1, \eta) + \ldots + \pi_k f(x; \phi_k, \eta)$$

- $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)$ are *mixture proportions*
- $\boldsymbol{\phi} = (\phi_1, ..., \phi_k)$ are (possibly vector) *component-specific parameters*
  - ϕ_j belongs to component j
- η is a (possibly vector) *common parameter* to all components
- f is a density
- Denote $\theta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$

The Label Switching problem

- For any permutation ν of 1,. . . ,k, define the corresponding permutation of θ as:

$$\nu(\theta) = \nu(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = ((\pi_{\nu(1)}, ..., \pi_{\nu(k)}), (\phi_{\nu(1)}, ..., \phi_{\nu(k)}), \eta)$$

- The problem is that the likelihood

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} \{\pi_1 f(x_i; \phi_1, \eta) + ... + \pi_k f(x_i; \phi_k, \eta)\}$$

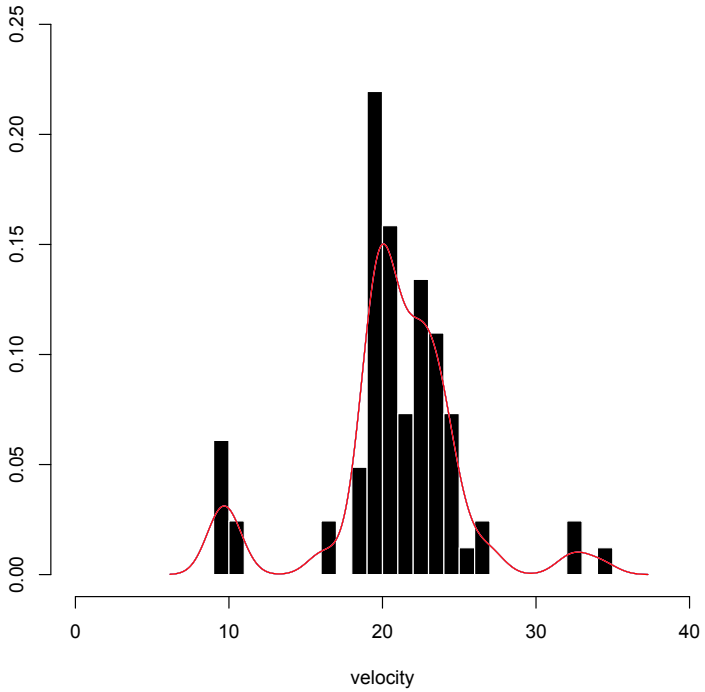
  is the same for all permutations of θ
- If $p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ is the same for all permutations of θ, posterior distribution will be symmetric
- Issues occur when estimating quantities relating to individual mixtures

Overview

Example - Galaxy Data

- 82 data points consisting of velocities of distant galaxies diverging from our own, from six conic sections of the *corona borealis*

$$p(x|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 \mathcal{N}(x; \mu_1, \sigma_1^2) + ... + \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2)$$



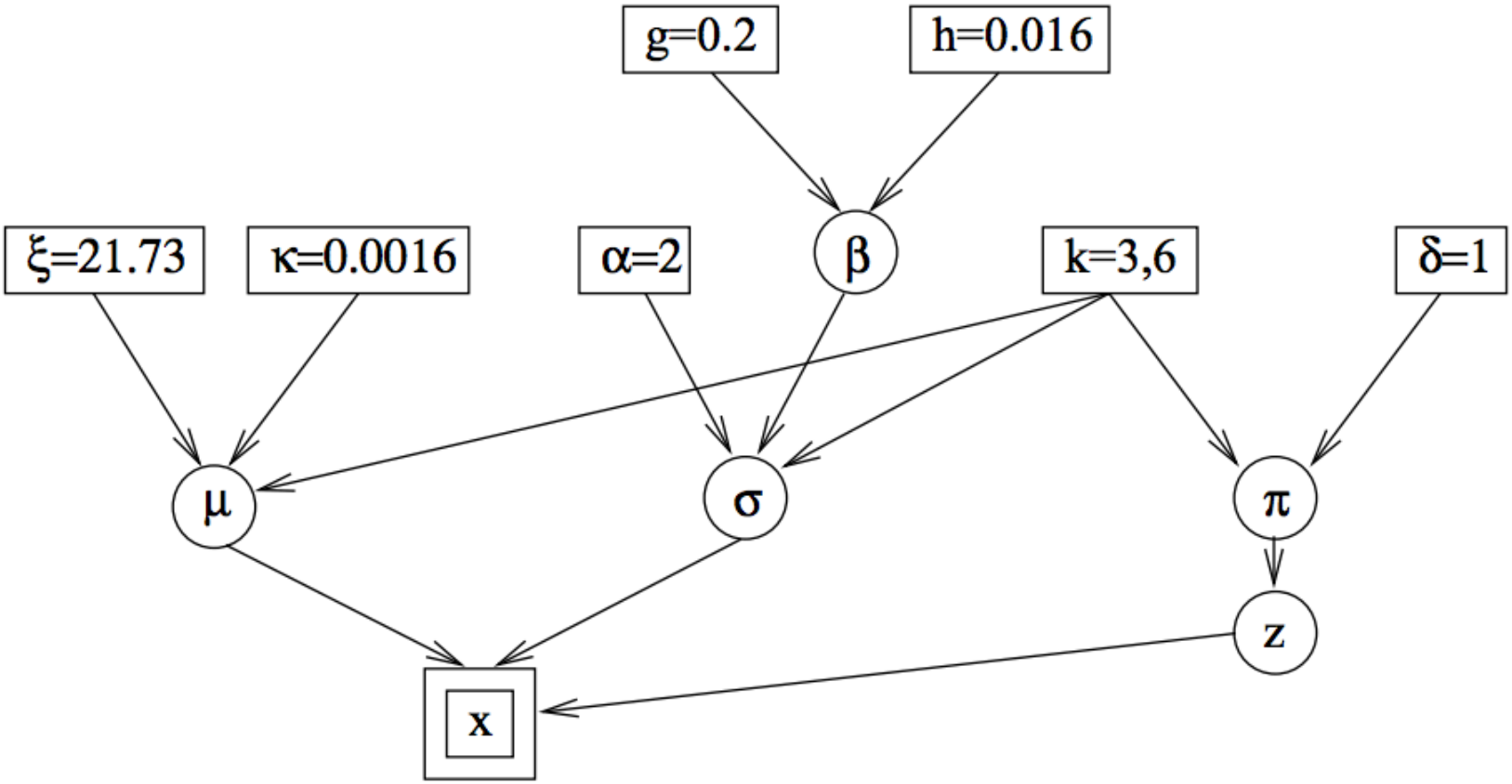
Random β Model - Initial

- Priors
  - $w \sim \mathcal{D}(\delta, ..., \delta)$
    - $\delta = 1$
  - $\mu_j \sim N(\xi, \kappa^{-1})$
    - $\xi = 21.73$ (midpoint of interval)
    - $R = 25.11$ (length of interval)
    - $\kappa = 1/R^2 = 0.0016$
  - $\sigma_j^{-2} \sim \Gamma(\alpha, \beta)$
    - $\alpha = 2$
  - $\beta \sim \Gamma(g, h)$
    - $g = 0.2$
    - $h = 10/R^2 = 0.016$

Random β Model - Gibbs Sampling

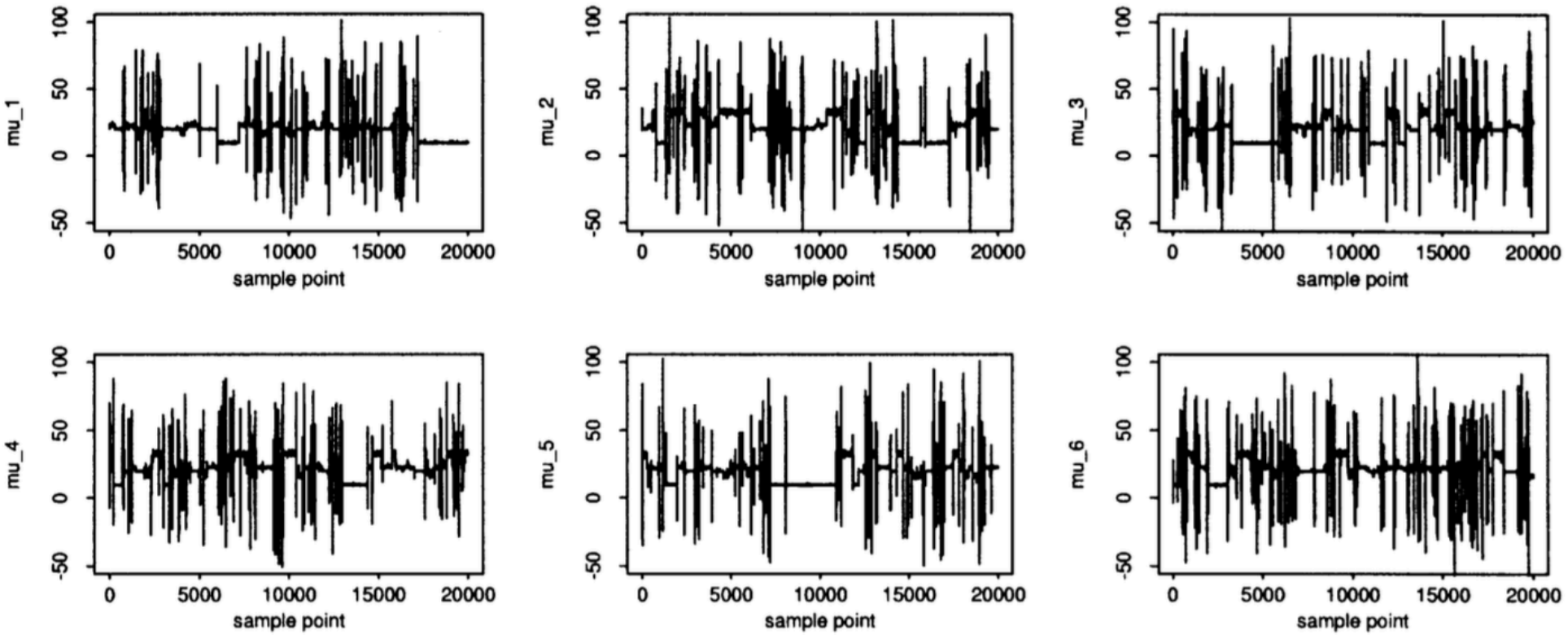
- Semi-Conjugate Priors
  - $w|\cdots \sim \mathcal{D}(\delta + n_1, ..., \delta + n_k)$
    - $n_j =$ number of observations in component j
  - $p(z_i = j|\dots) \propto \frac{w_j}{\sigma_j} \exp\left(-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right)$
    - $z_i =$ component x_i believed to belong to
  - $\mu_j|\cdots \sim N\left(\frac{\sigma_j^{-2}\sum_{i:z_i=j} x_i + \kappa\xi}{\sigma_j^{-2} n_j + k}, (\sigma_j^{-2} n_j + \kappa)^{-1}\right)$
  - $\sigma_j^{-2}|\cdots \sim \Gamma(\alpha + \frac{1}{2} n_j, \beta + \frac{1}{2}\sum_{i:z_i=j}(x_i - \mu_j)^2)$
  - $\beta|\cdots \sim \Gamma(g + \kappa\alpha, h + \sum_j \sigma_j^{-2})$

Hierarchical Model

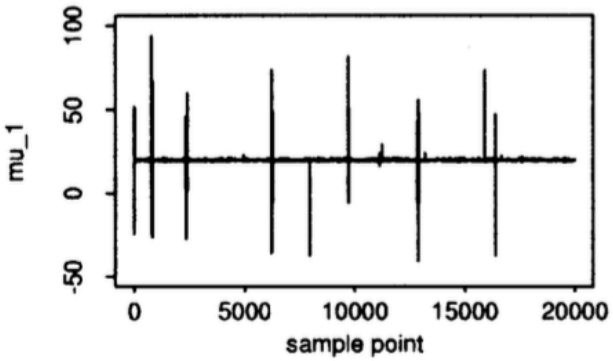


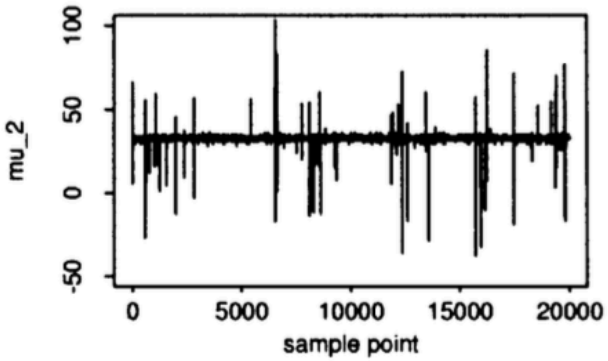
Gibbs Output

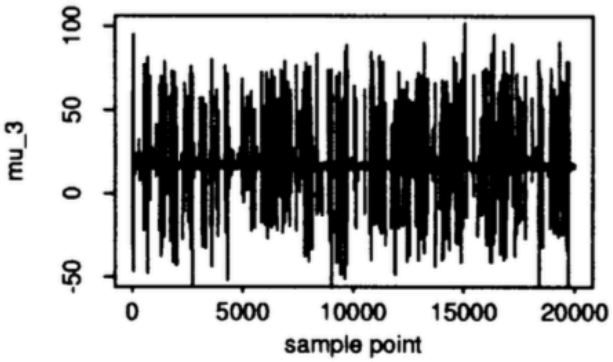
- The means each take values from each region of the components

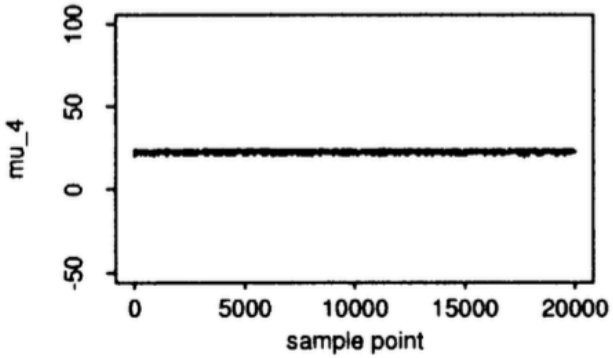


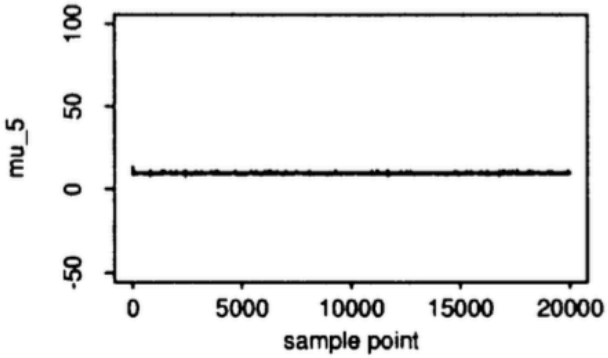
After Relabelling

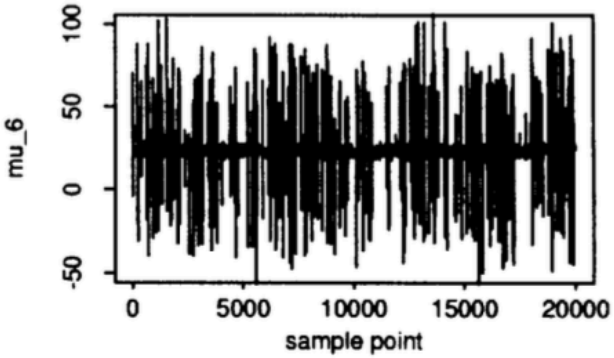












Overview

Past Approaches

- Artificial identifiability constraints
  - satisfied by only one permutation of θ for each θ
  - i.e.

$$\pi_1 < \pi_2 < \ldots < \pi_k$$
$$\mu_1 < \mu_2 < \ldots < \mu_k$$
$$\vdots$$

  - breaks symmetry of the prior, and thus the posterior
  - however, with so many constraint options, label switching remains
- Celeux *et al.* (1996) recommend three methods, but require knowing the 'true' values of the parameters, making the process difficult
- Richardson and Green (1997) suggest that the MCMC output should be post-processed
- New methods use a more decision theoretic approach, post-process, attempting to minimize the posterior expectation of some suggested loss functions

Relabelling Algorithms

- Post-process
- Choosing a single action a from a set of possible actions $\mathcal{A}$
- Decision Theoretic Approach
  - define a loss function $\mathcal{L} : \mathcal{A} \times \Theta \to R$
  - $\mathcal{L}(a; \theta) =$ loss from choosing action a when true parameter value is θ
  - choose action $\hat{a}$ that minimizes the posterior expected loss (or *risk*):

$$\mathcal{R}(a) = E\{\mathcal{L}(a; \theta)|\boldsymbol{x}\}$$

- Following the invariance of the likelihood, use loss function of the form:

$$\mathcal{L}(a; \theta) = \min_{\nu} [\mathcal{L}_0\{a : \nu(\theta)\}]$$

for some $\mathcal{L}_0 : \mathcal{A} \times \Theta \to R$

Relabelling Algorithms cont.

- $\theta^{(1)}, ..., \theta^{(N)}$ from sampled states from a Markov chain, with distribution $p(\theta|\boldsymbol{x})$, then approximate the risk $\mathcal{R}(a)$ by:

$$\tilde{\mathcal{R}}(a) = \frac{1}{N}\sum_{t=1}^{N} \min_{\nu_t}[\mathcal{L}_0\{a; \nu_t(\theta^{(t)})\}] = \min_{\nu_1,\ldots,\nu_N}\left[\frac{1}{N}\sum_{t=1}^{N}\mathcal{L}_0\{a; \nu_t(\theta^{(t)})\}\right],$$

  and choose $\hat{a}$ to minimize $\tilde{\mathcal{R}}$

- Stephens describes two algorithms incorporating the decision theoretic approach; we will focus on the second

Algorithm 2

- Suppose we want to cluster the observations into k groups
- Create matrix $Q = (q_{ij})$, and $P(\theta) = (p_{ij}(\theta))$
- Start with initial values for $\nu_1, ..., \nu_N$; iterate until convergence:

Step 1: choose $\hat{Q} = (\hat{q}_{ij})$ to minimize

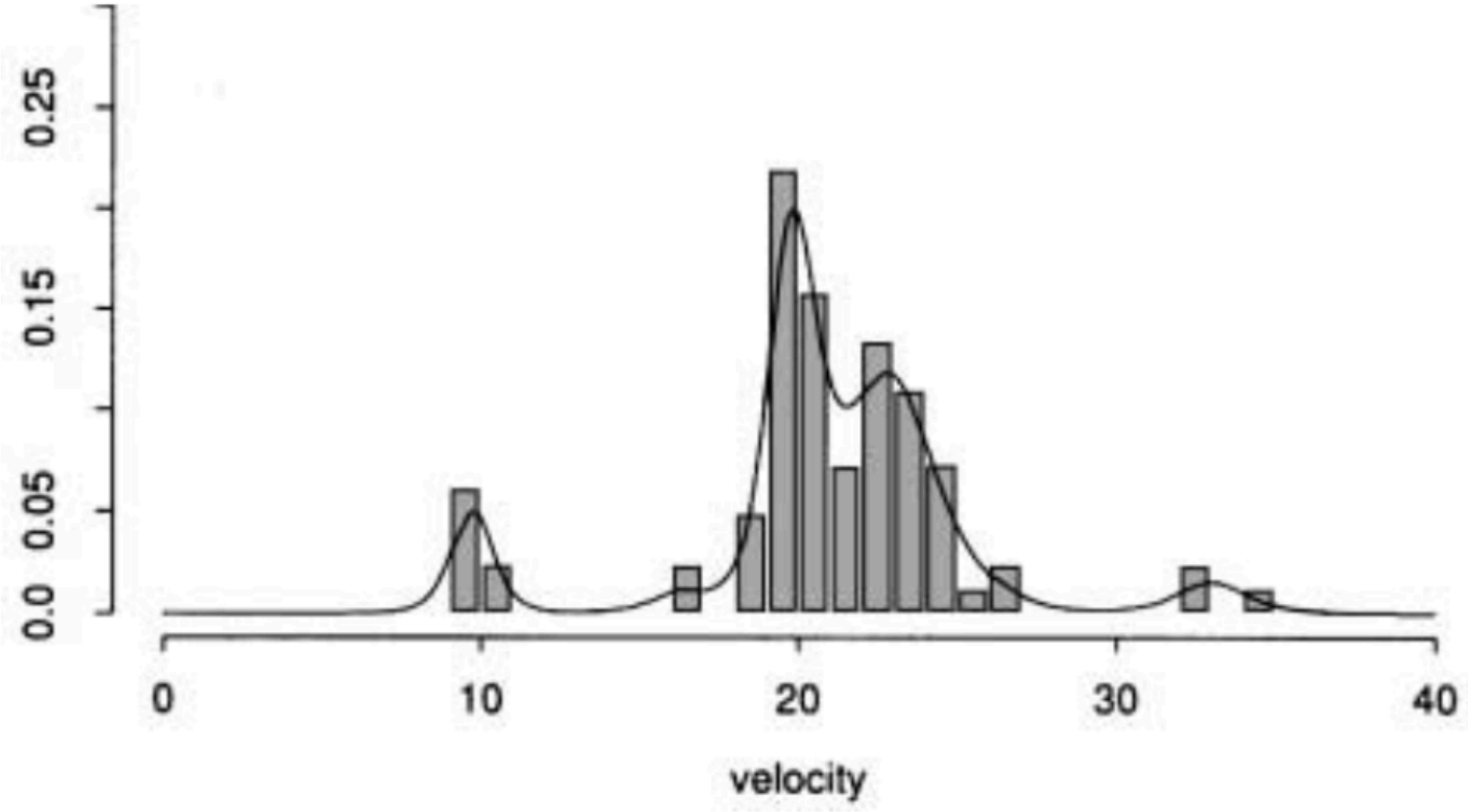
$$\sum_{t=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\{\nu_t(\theta^{(t)})\} log \left[\frac{p_{ij}\{\nu_t(\theta^{(t)})\}}{\hat{q}_{ij}} \right]$$

Step 2: for $t = 1, ..., N$ choose ν_t to minimize

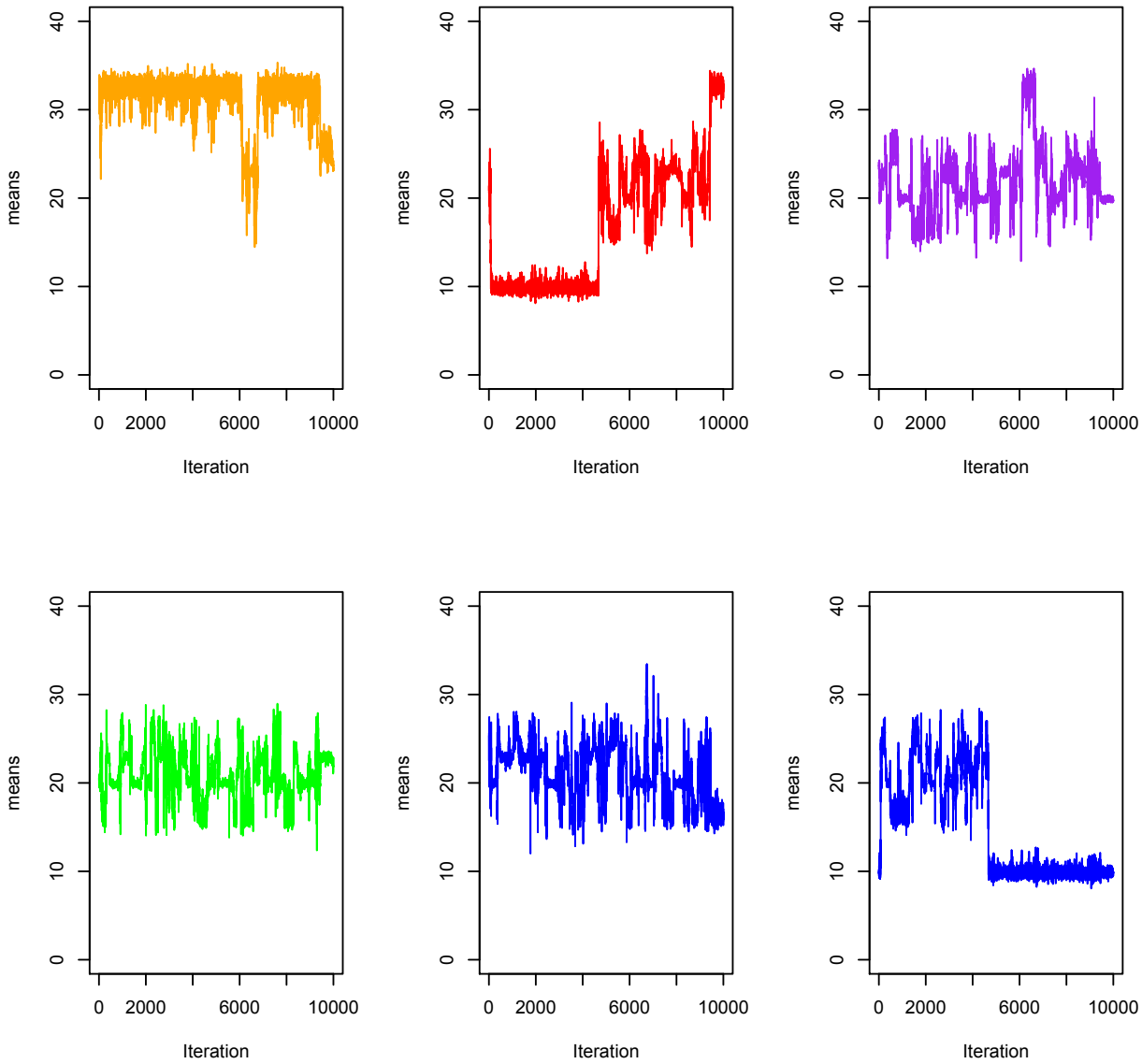
$$\sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\{\nu_t(\theta^{(t)})\} log \left[\frac{p_{ij}\{\nu_t(\theta^{(t)})\}}{\hat{q}_{ij}} \right]$$

Overview

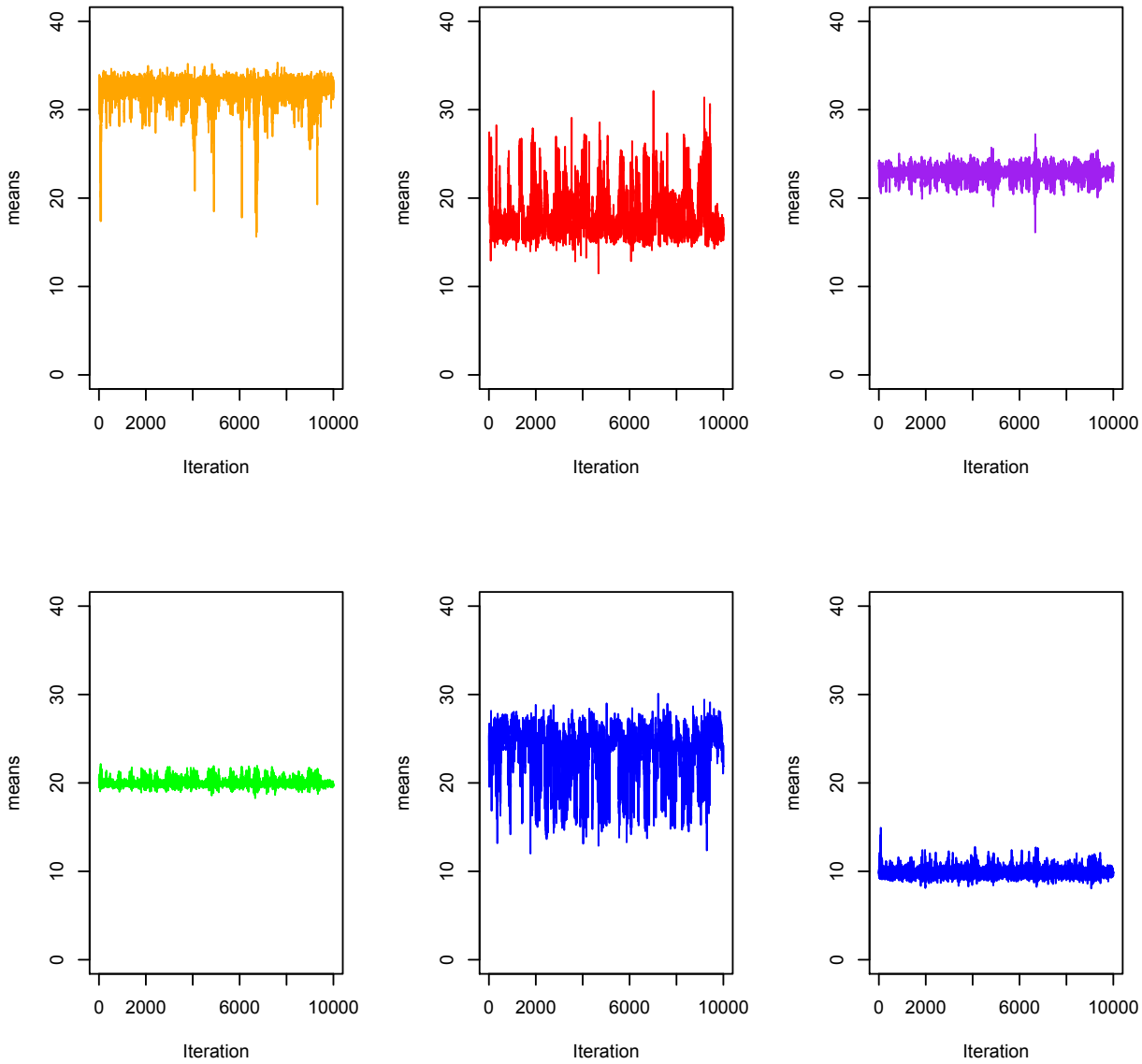
Reminder of Galaxy data



Initial MCMC - Gravity

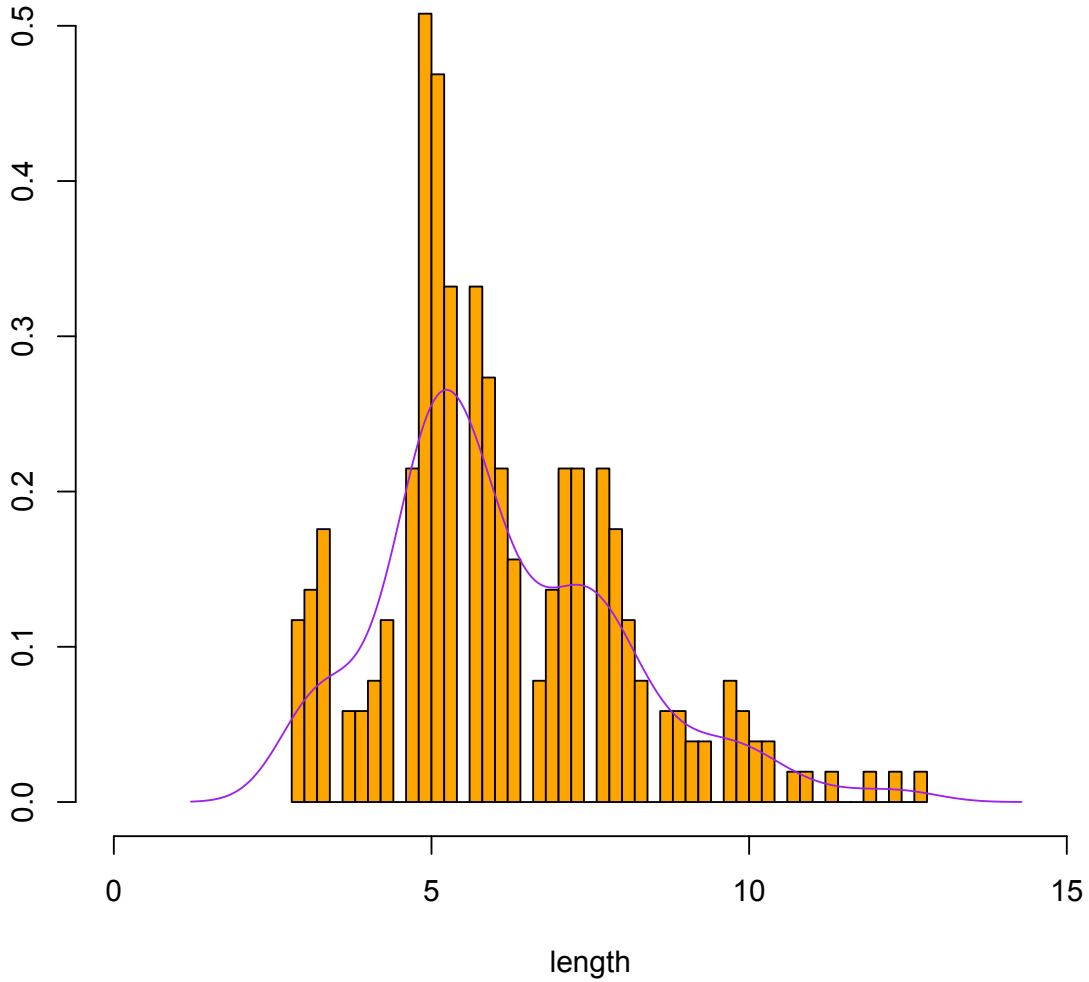


Stephens's Method Applied

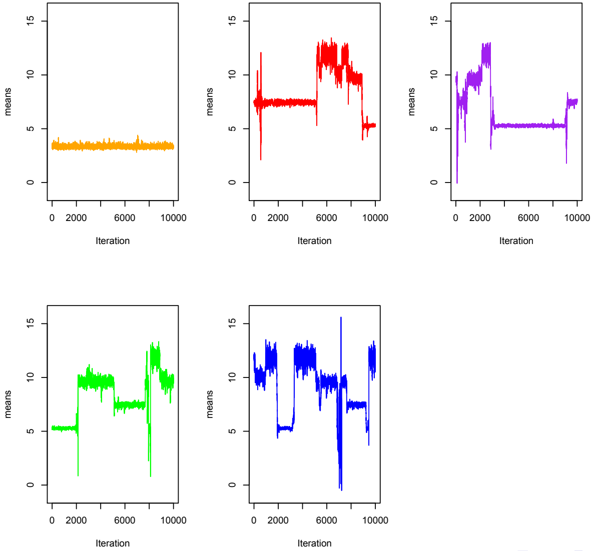


Example - Fish data

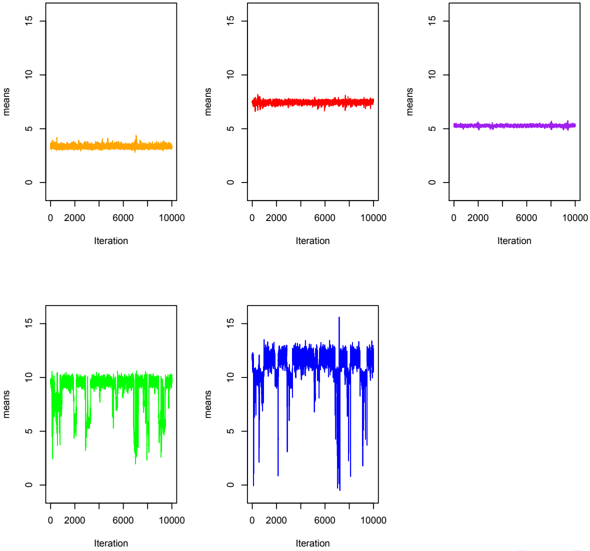
- Data consisting of 256 snapper length measurements



Initial MCMC - Fish



Stephens's Method Applied



Conclusion

- Not perfect
- Many options available:
  - STEPHENS
  - PRA (Pivotal Reordering Algorithm)
  - ECR (Equivalence Classes Representatives)
  - SJW (Probabilistic Relabelling Algorithm)
  - AIC (Artificial Identifiability Constraints)
  - *etc.*
- Compare all outputs, and choose the algorithm that fits the best

References I

Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

Zhihui Liu et al. *Bayesian Mixture Models*. PhD thesis, 2010.

Panagiotis Papastamoulis. label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*, 2015.

Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

Kathryn Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

Matthew Stephens. Bayesian methods for mixtures of normal distributions, 1997.

References II

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1985.

Thank you

Questions?