Multivariate Nonparametric Mixture Models

Xiaotian Zhu[1] and David R. Hunter[2]

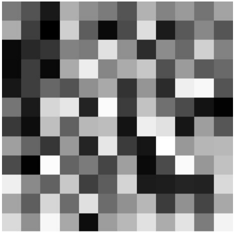
[1]Google
Pittsburgh, PA

[2]Department of Statistics
Penn State University

PSU Astrostatistics Lunch, March 30, 2023

A Motivating Example: Classifying Image Excerpts

- **Goal:** Extracting intrinsic structure in images by clustering and finding complete set of efficient linear basis functions.

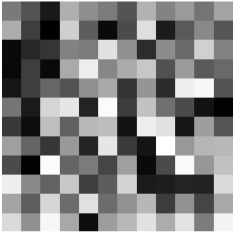


painting?

newspaper?

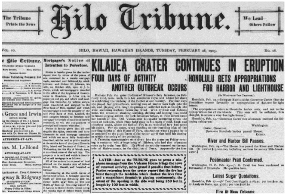
A Motivating Example: Classifying Image Excerpts

- **Goal:** Extracting intrinsic structure in images by clustering and finding complete set of efficient linear basis functions.



painting?

newspaper?



Outline

1. Non- and semi-parametric mixture models; the identifiability problem
2. An EM-like framework for estimation
3. Some multivariate clustering problems and identifiability
4. Combining NP mixture models with Independent Components Analysis

Outline

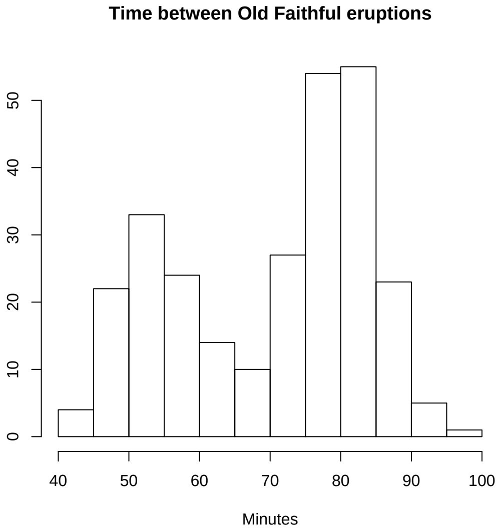
Let us first introduce nonparametric finite mixtures

$$X \sim \underbrace{g(x)}_{\substack{\text{mixture}\\ \text{density}}} = \int \underbrace{f_\phi(x)}_{\substack{\text{component}\\ \text{density}}} \underbrace{dQ(\phi)}_{\substack{\text{mixing}\\ \text{distribution}}} \tag{1}$$

- Sometimes, $Q(\cdot)$ is the "nonparametric" part;
  e.g., work by Bruce Lindsay assumes $Q(\cdot)$ is unrestricted.
- However, in this talk we assume that
  - $f_\phi(\cdot)$ is (mostly) unrestricted
  - $Q(\cdot)$ has finite support

So (1) becomes

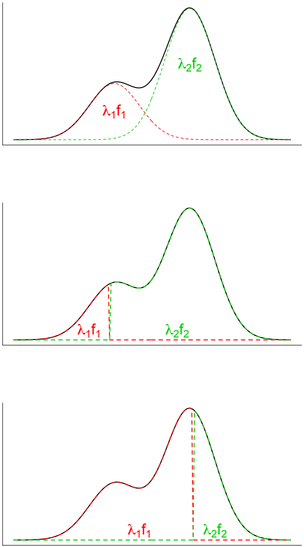
$$g(x) = \sum_{j=1}^{m} \lambda_j f_j(x) \qquad \ldots \text{and we assume } m \text{ is known.}$$



from www.nps.gov/yell

- Let $m = 2$, so assume we have a sample from
$$\lambda_1 f_1(x) + \lambda_2 f_2(x).$$
- Why do we need any assumptions on f_j?

With no assumptions, parameters are not identifiable



- Multiple different parameter combinations

$$(\lambda_1, \lambda_2, f_1, f_2)$$

  give the same mixture density.
- Thus, some constraints on f_j are necessary.
- NB: Sometimes, there is no obvious multi-modality.

The univariate case is identifiable under some assumptions

It is possible to show[1] that if

$$g(x) = \sum_{j=1}^{2} \lambda_j f_j(x),$$

the λ_j and f_j are uniquely identifiable from g if $\lambda_1 \neq 1/2$ and

$$f_j(x) \equiv f(x - \mu_j)$$

for some density $f(\cdot)$ that is *symmetric about the origin*.

[1] cf. Bordes, Mottelet, and Vandekerkhove (2006); Hunter, Wang, and Hettmansperger (2007)

Outline

Here's a review of EM-ology for mixture models:

EM preliminaries: A "complete" observation $(X, \mathbf{Z})$ consists of:

- The "observed" data X
- The "unobserved" vector $\mathbf{Z}$, defined by

$$\text{for } 1 \le j \le m,\ Z_j = \begin{cases} 1 & \text{if } X \text{ comes from component } j \\ 0 & \text{otherwise} \end{cases}$$

Standard EM for finite mixtures looks like this:

- **E-step:** Amounts to finding the conditional expectation of each $\mathbf{Z}_i$:

$$\hat{Z}_{ij} \stackrel{\text{def}}{=} \mathrm{E}_{\hat{\boldsymbol{\theta}}}(Z_{ij} \mid \mathbf{X} = \mathbf{x}) = P_{\hat{\boldsymbol{\theta}}}(Z_{ij} = 1 \mid \mathbf{X} = \mathbf{x}) = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\hat{\boldsymbol{\lambda}} \cdot \hat{\mathbf{f}}(x_i)}$$

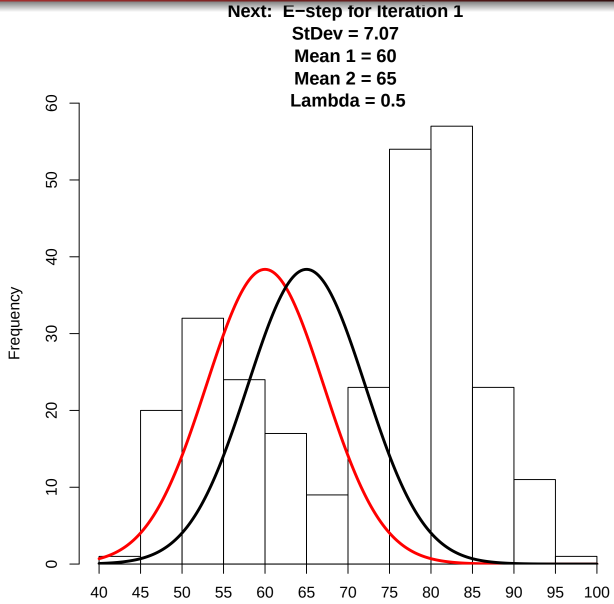
- **M-step:** Amounts to maximizing the "expected complete data loglikelihood"

$$L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{Z}_{ij} \log \left[\lambda_j f_j(x_i)\right] \quad \Longrightarrow \quad \hat{\boldsymbol{\lambda}}^{\text{next}} = \frac{1}{n} \sum_i \hat{\mathbf{Z}}_i$$

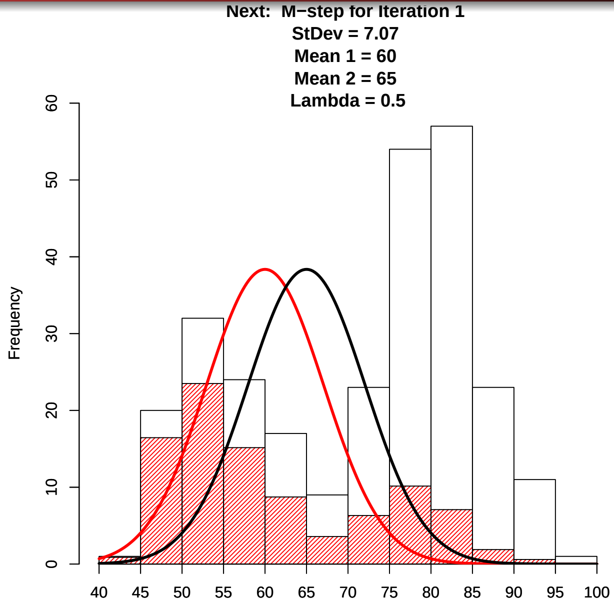
- **Iterate:** Let $\hat{\boldsymbol{\theta}}^{\text{next}} = \arg\max_{\theta} L_c(\boldsymbol{\theta})$ and repeat.

N.B.: Usually, $f_j(x) \equiv f(x; \phi_j)$. *We let* $\boldsymbol{\theta}$ *denote* $(\boldsymbol{\lambda}, \phi)$.

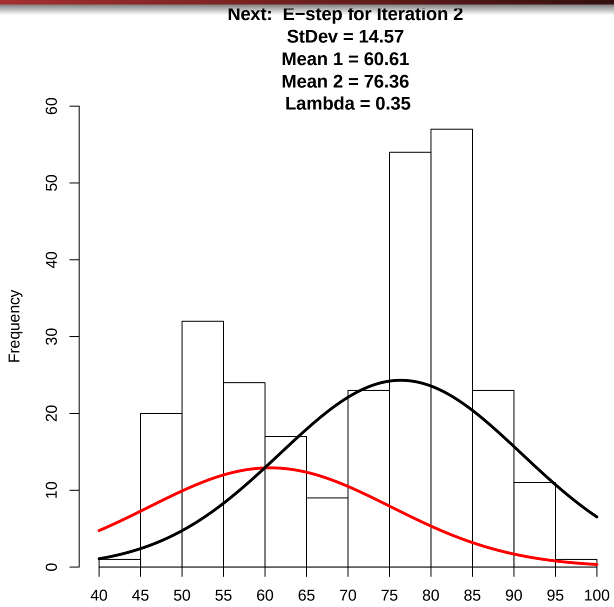
EM movie based on Old Faithful Data



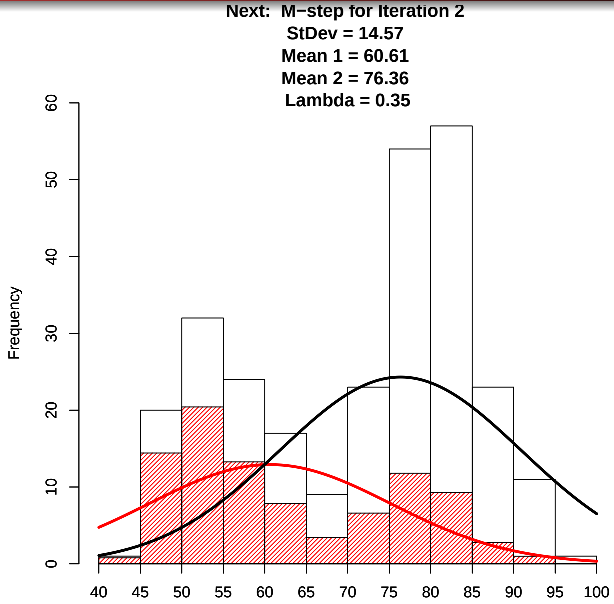
EM movie based on Old Faithful Data



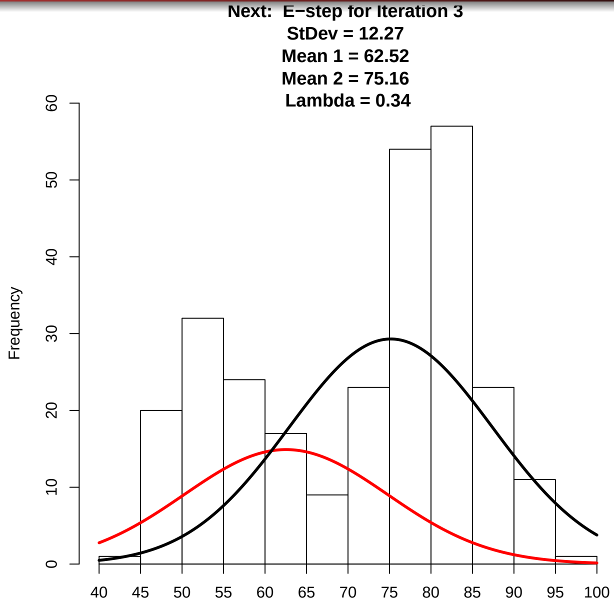
EM movie based on Old Faithful Data



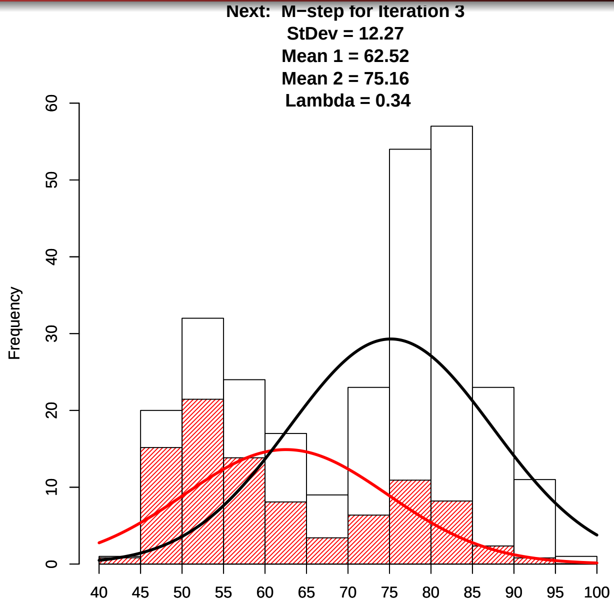
EM movie based on Old Faithful Data



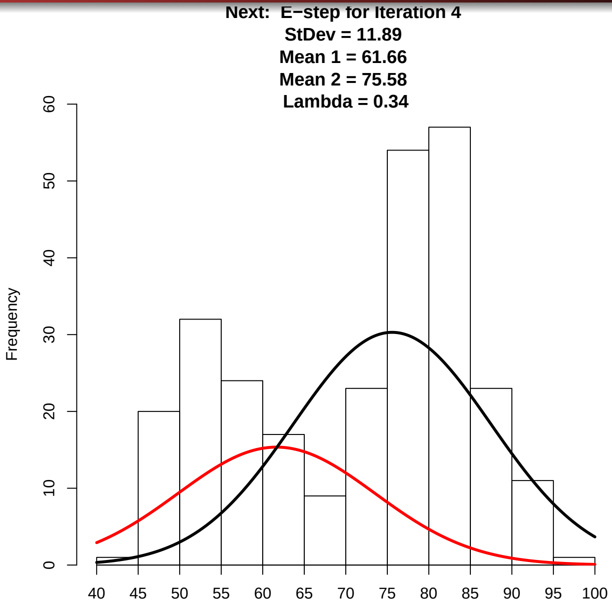
EM movie based on Old Faithful Data



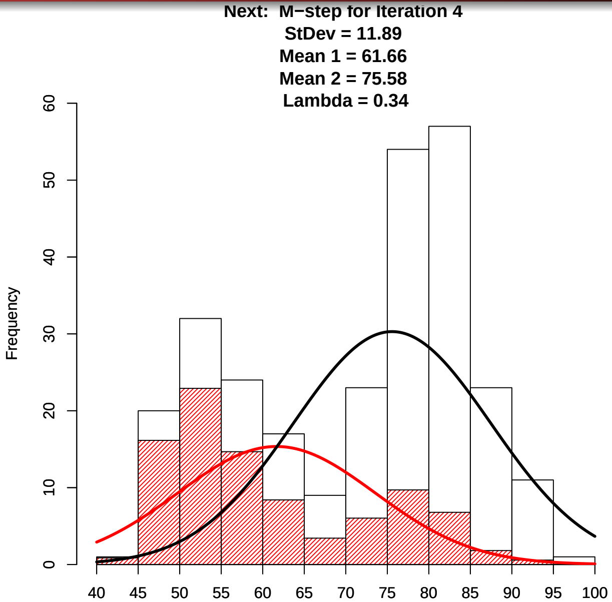
EM movie based on Old Faithful Data



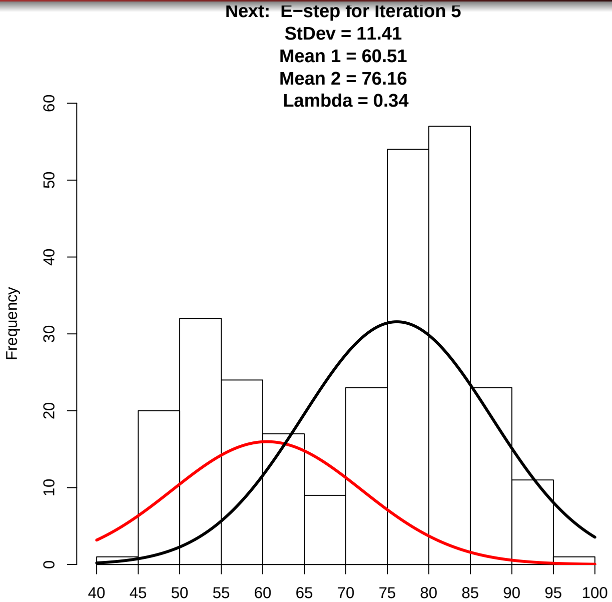
EM movie based on Old Faithful Data



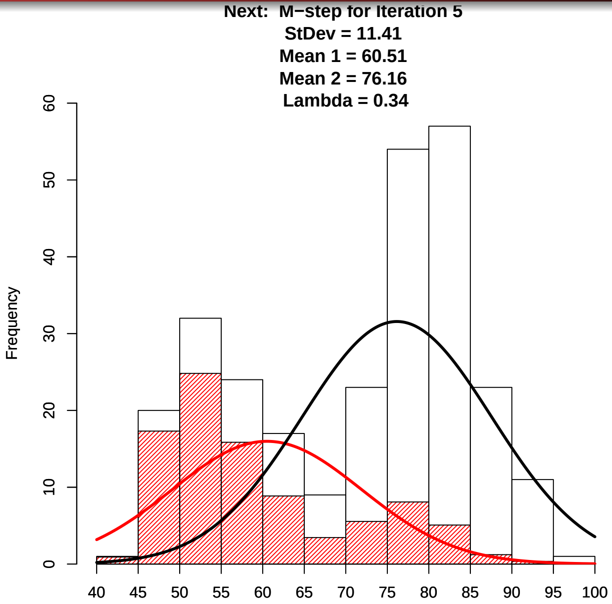
EM movie based on Old Faithful Data



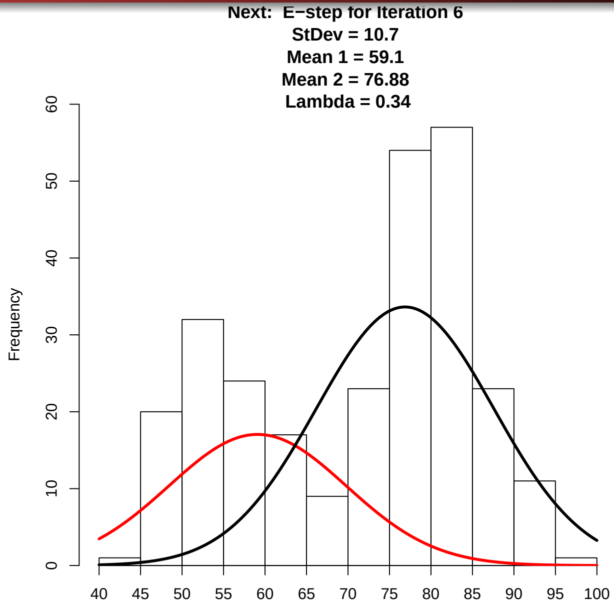
EM movie based on Old Faithful Data



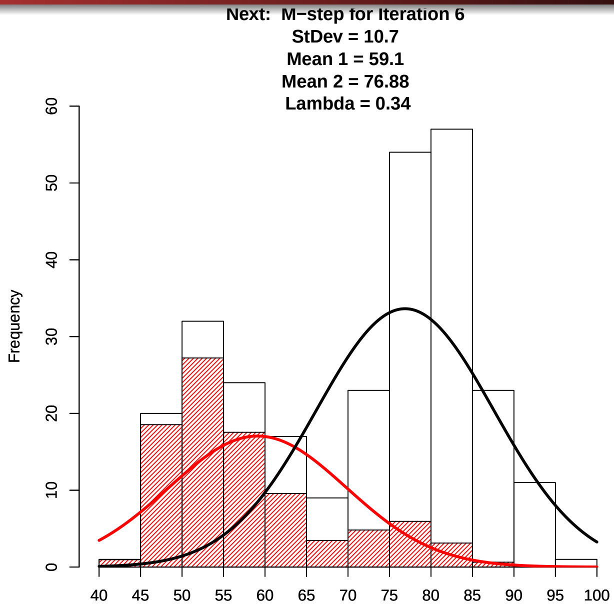
EM movie based on Old Faithful Data



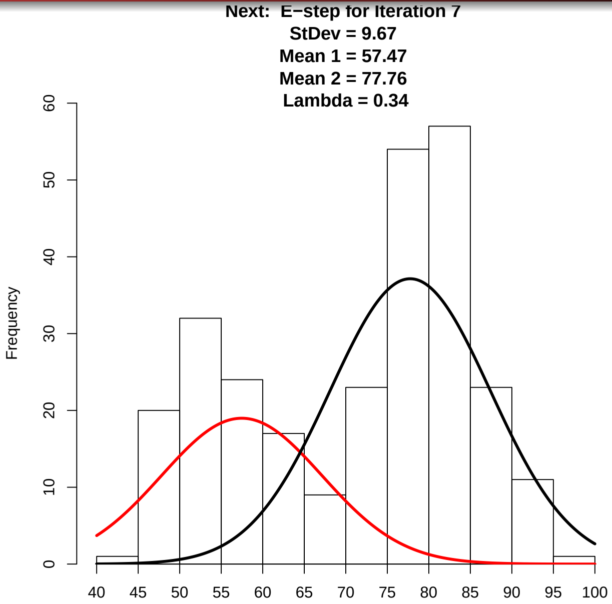
EM movie based on Old Faithful Data



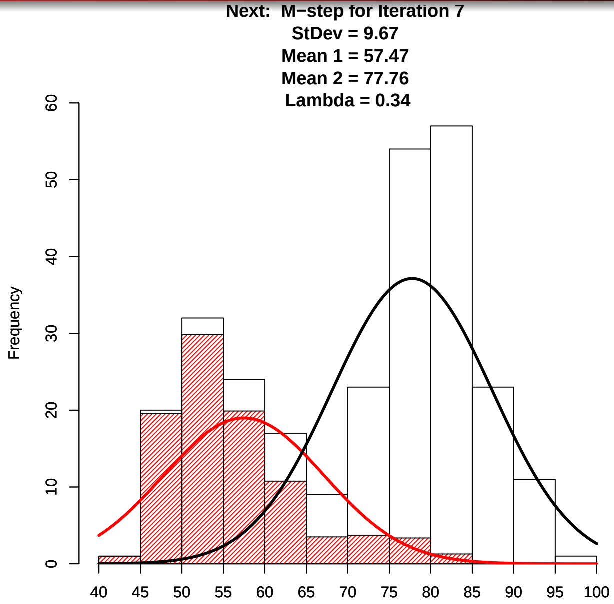
EM movie based on Old Faithful Data



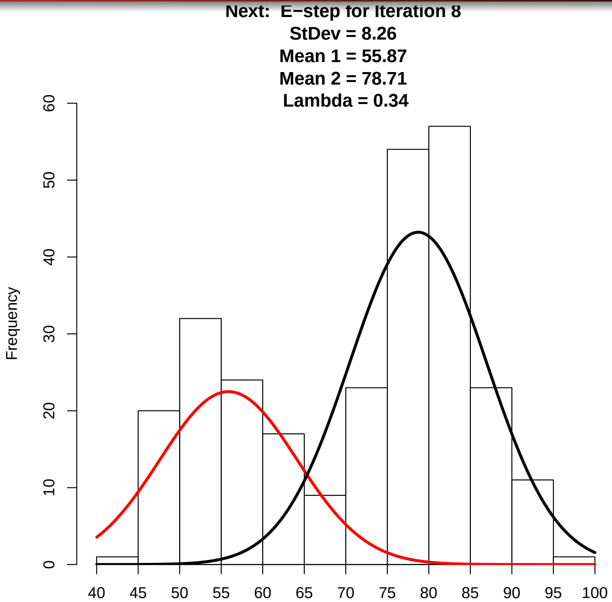
EM movie based on Old Faithful Data



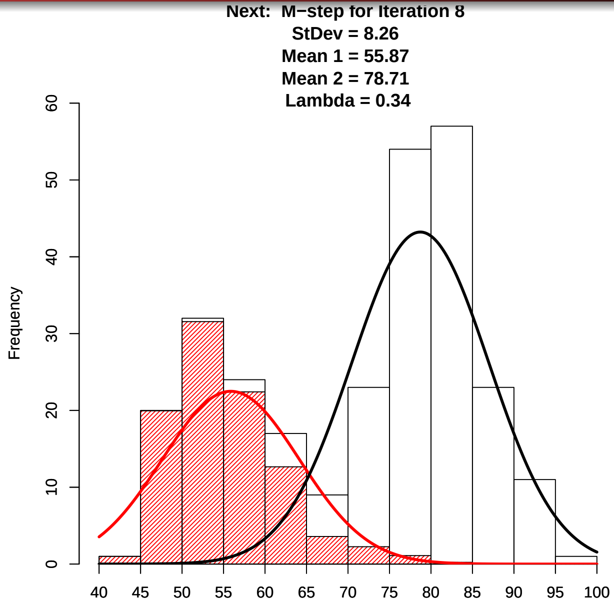
EM movie based on Old Faithful Data



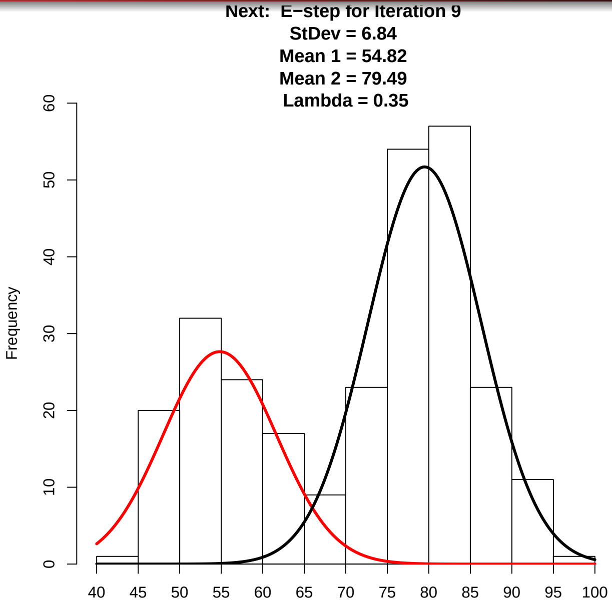
EM movie based on Old Faithful Data



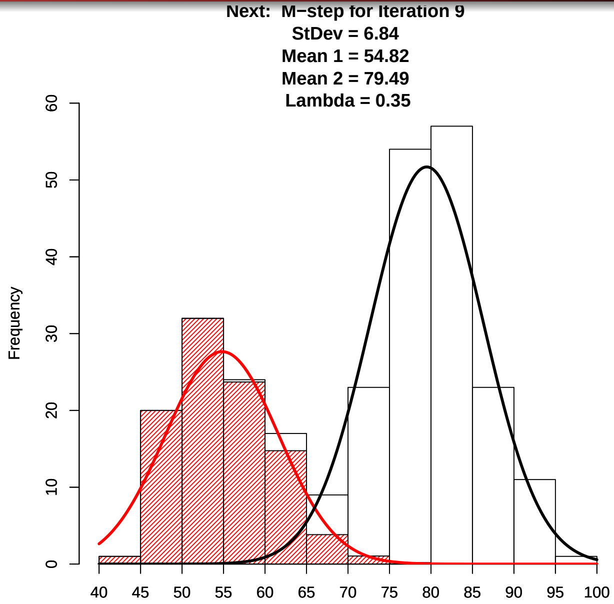
EM movie based on Old Faithful Data



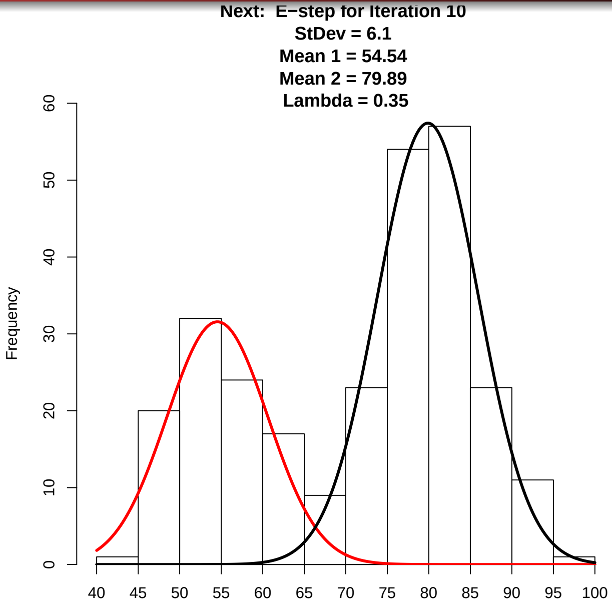
EM movie based on Old Faithful Data



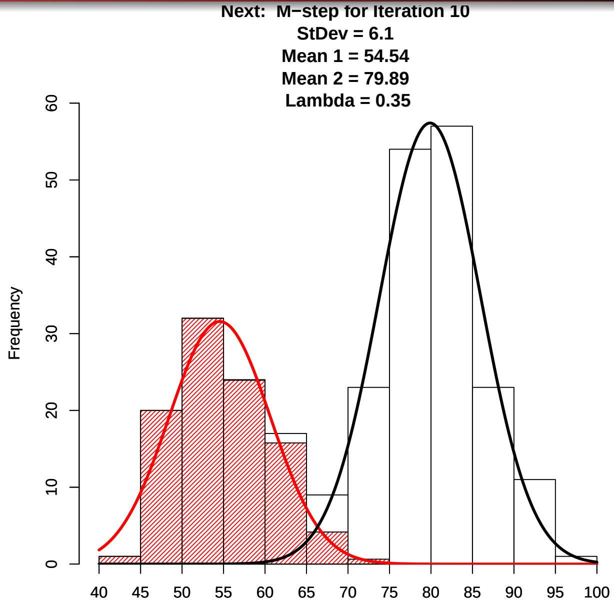
EM movie based on Old Faithful Data



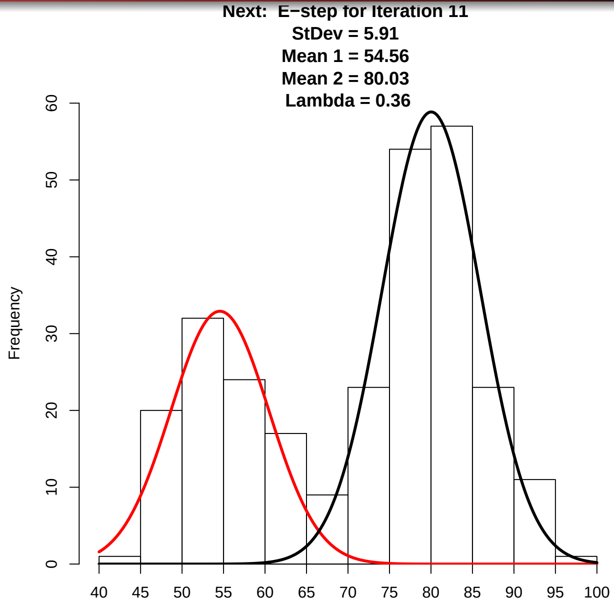
EM movie based on Old Faithful Data



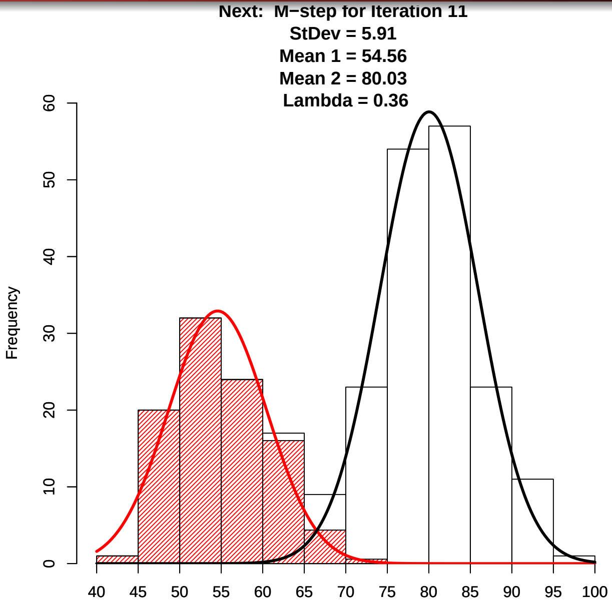
EM movie based on Old Faithful Data



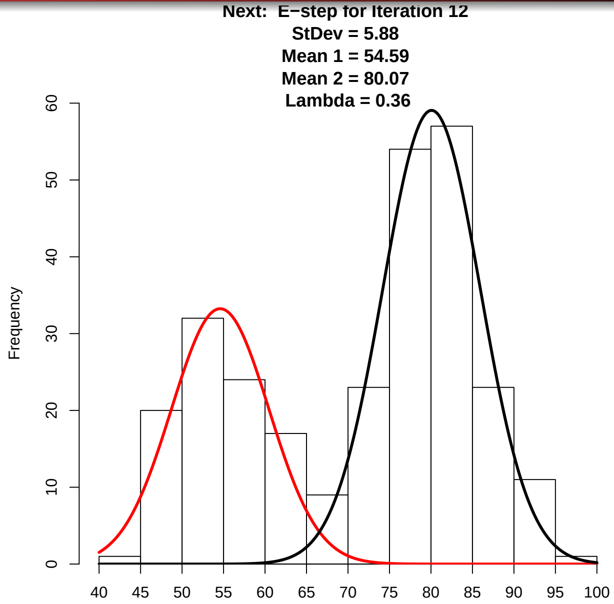
EM movie based on Old Faithful Data



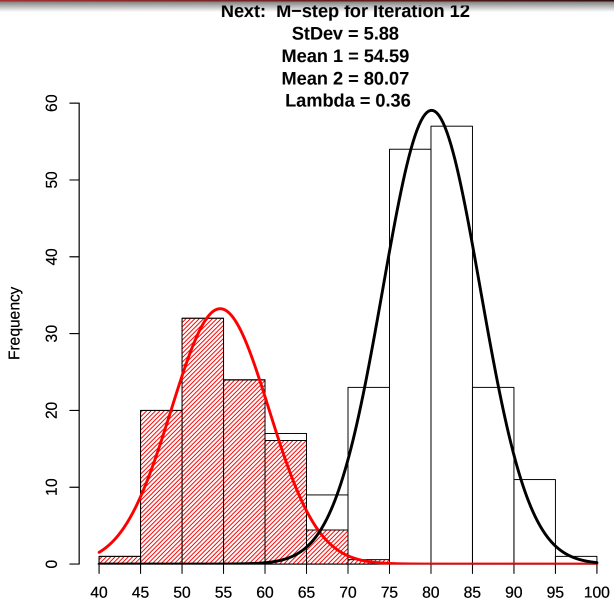
EM movie based on Old Faithful Data



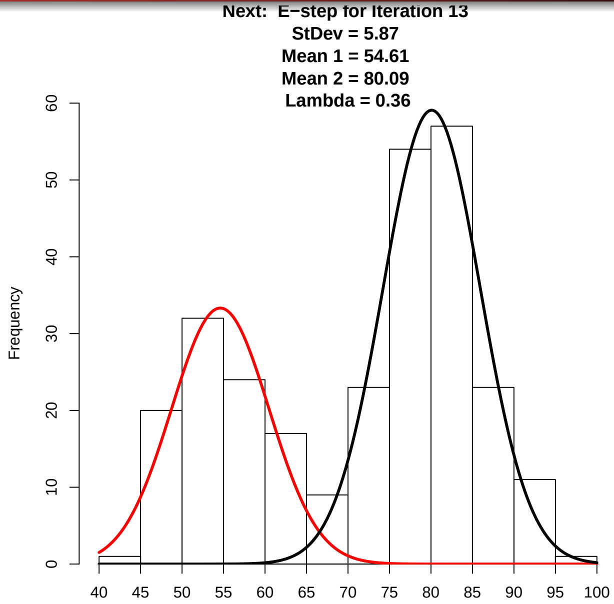
EM movie based on Old Faithful Data



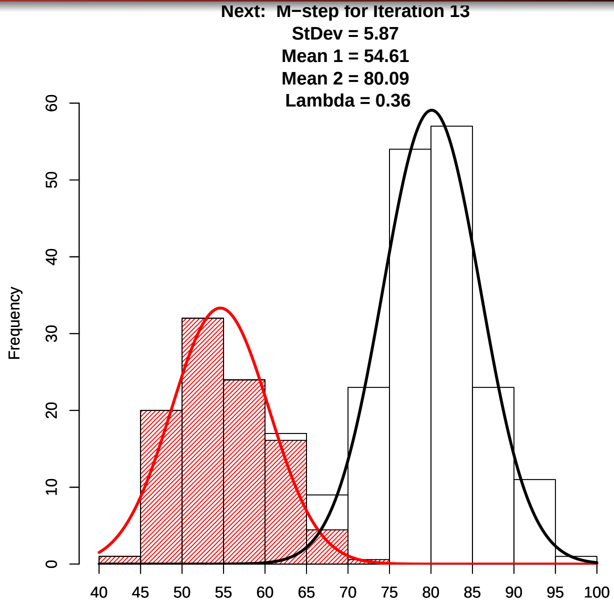
EM movie based on Old Faithful Data



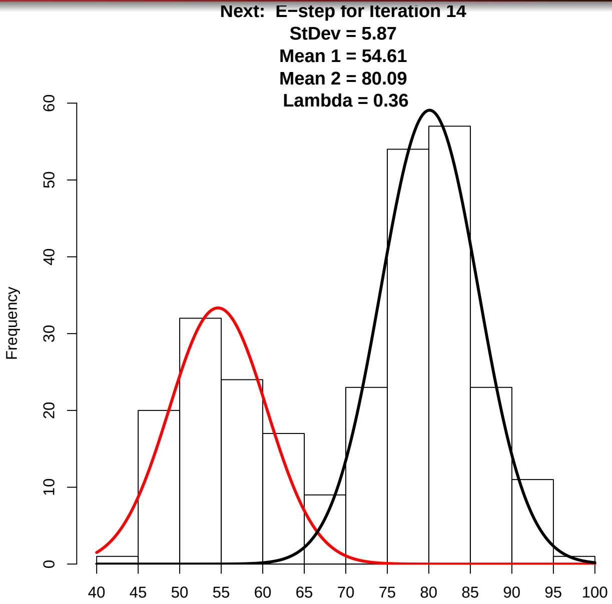
EM movie based on Old Faithful Data



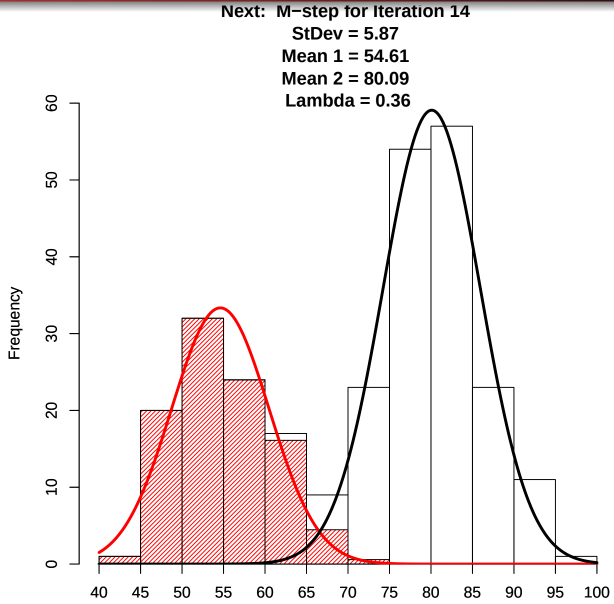
EM movie based on Old Faithful Data



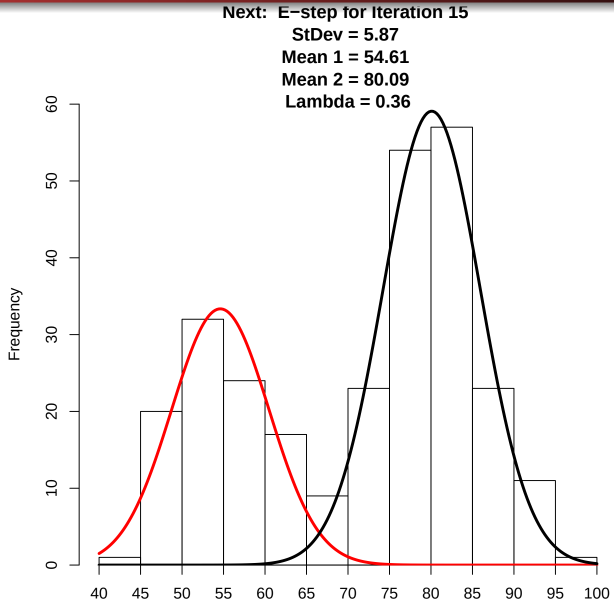
EM movie based on Old Faithful Data



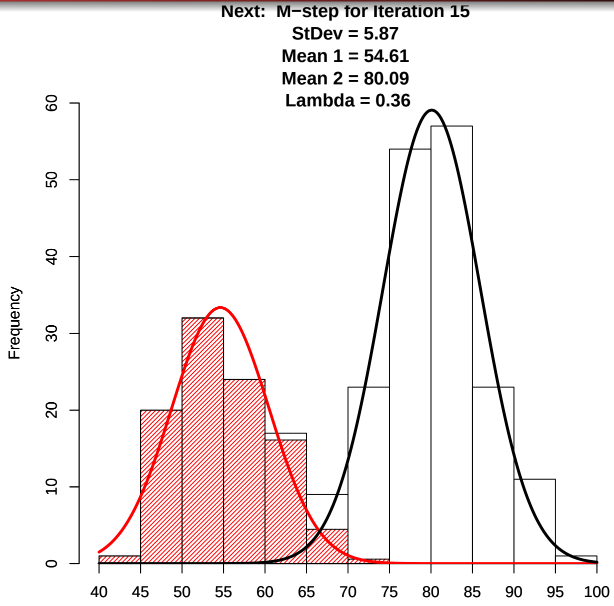
EM movie based on Old Faithful Data



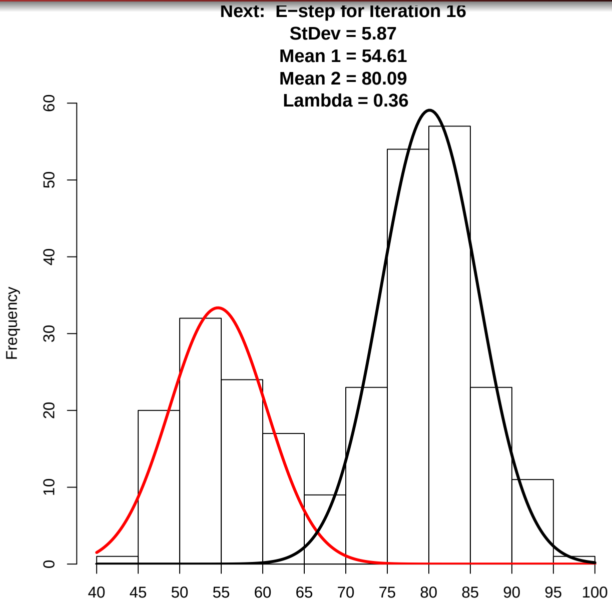
EM movie based on Old Faithful Data



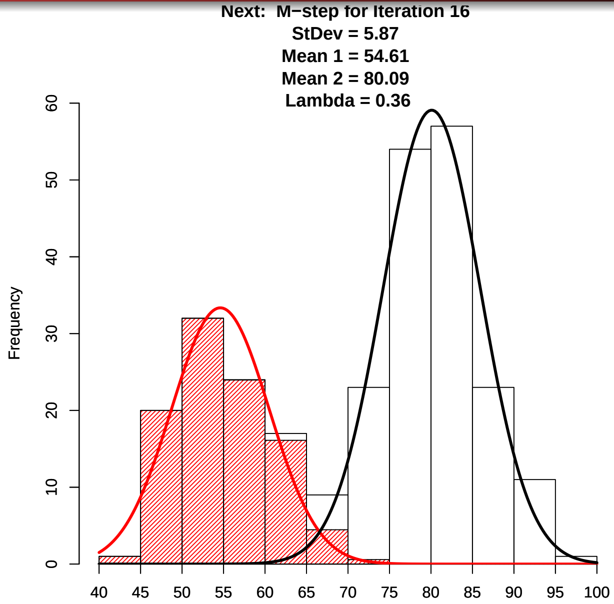
EM movie based on Old Faithful Data



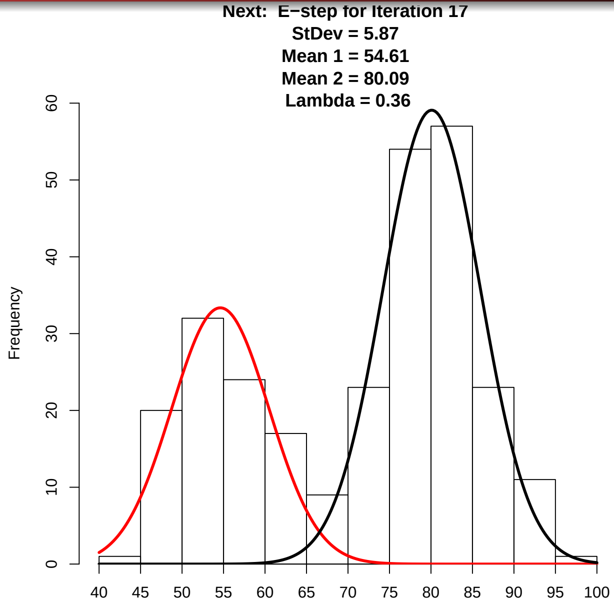
EM movie based on Old Faithful Data



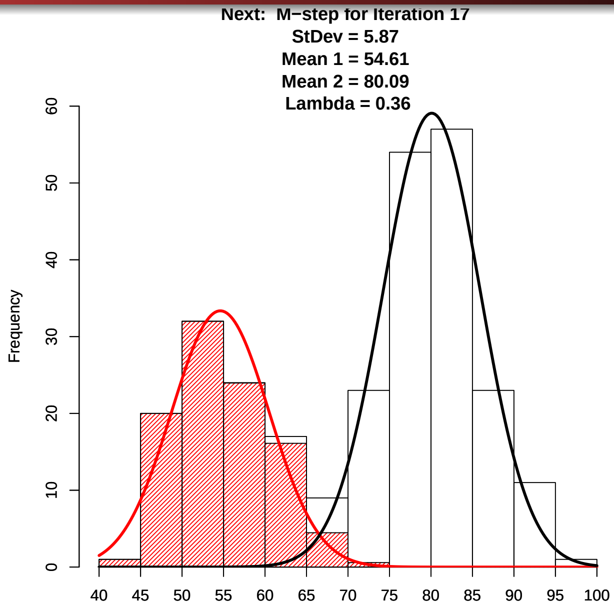
EM movie based on Old Faithful Data



EM movie based on Old Faithful Data



EM movie based on Old Faithful Data



Recall the non- (or semi-)parametric problem

We believe we have an i.i.d. sample from

$$g(x) = \sum_{j=1}^{2} \lambda_j f_j(x),$$

where we assume $\lambda_1 \neq 1/2$ and

$$f_j(x) \equiv f(x - \mu_j)$$

for some density $f(\cdot)$ that is *symmetric about the origin*.

Thus, the parameters to estimate are f_j, λ_j, and μ_j for $j = 1$ and $j = 2$.

We may modify the usual EM algorithm

E-step: Same as usual:

$$\hat{Z}_{ij} \equiv \mathrm{E}_{\hat{\boldsymbol{\theta}}}(Z_{ij} \,|\, \mathbf{X} = \mathbf{x}) = \frac{\hat{\lambda}_j \hat{f}(x_i - \hat{\mu}_j)}{\hat{\lambda}_1 \hat{f}(x_i - \hat{\mu}_1) + \hat{\lambda}_2 \hat{f}(x_i - \hat{\mu}_2)}$$

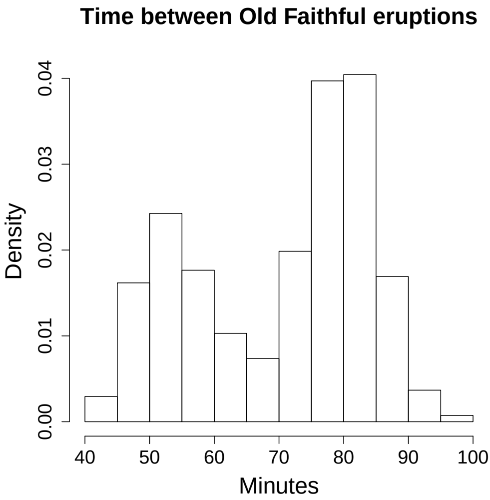
M-step: Maximize complete data "loglikelihood" for $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$:

$$\hat{\lambda}_j^{\text{next}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij} \qquad \hat{\mu}_j^{\text{next}} = (n\tilde{\lambda}_j)^{-1} \sum_{i=1}^{n} \hat{Z}_{ij} x_i$$

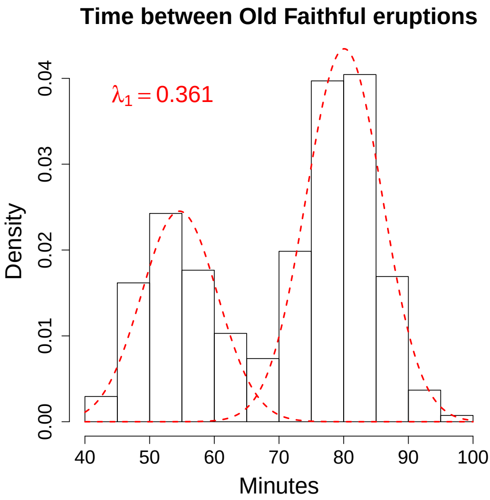
KDE-step: Update estimate of f (for some bandwidth h) by

$$\hat{f}^{\,\text{next}}(u) = (nh)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \hat{Z}_{ij} K\left(\frac{u - x_i + \hat{\mu}_j}{h}\right), \text{then symmetrize.}$$

Compare two solutions for Old Faithful data

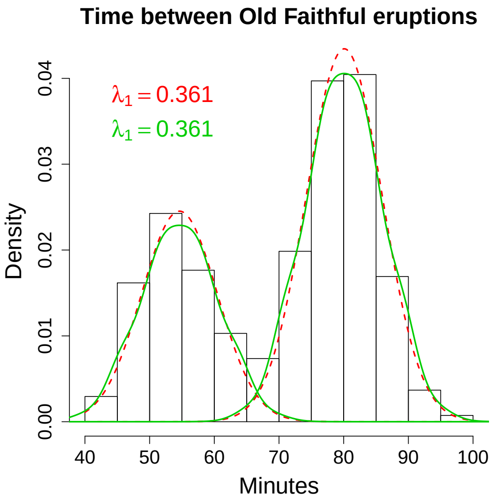


Compare two solutions for Old Faithful data



- Gaussian EM: $\hat{\mu} = (54.6, 80.1)$

Compare two solutions for Old Faithful data



- Gaussian EM: $\hat{\boldsymbol{\mu}} = (54.6, 80.1)$
- Semiparametric EM with bandwidth $= 4$: $\hat{\boldsymbol{\mu}} = (54.7, 79.8)$
- Both algorithms implemented in **mixtools** package for R (Benaglia, Chauveau, Hunter, & Young 2009).

Outline

Introducing the notion of conditional independence

Each density on $\mathbb{R}^r$ is assumed to be the product of its marginals:

$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_k)$$

Crucially, we do not assume a parametric form for f_{jk}.

- We call this model *conditional independence* (cf. Hall and Zhou, 2003; Qin and Leung, 2006)
- Very similar to repeated measures models:
  - In RM models, we often assume measurements are independent conditional on the individual.
  - Here, we have component-specific effects instead of individual-specific effects.
- CI aids interpretation; univariate densities easy to visualize

There exists an elegant identifiability result for conditional independence when $r \geq 3$

Recall the conditional independence finite mixture model:

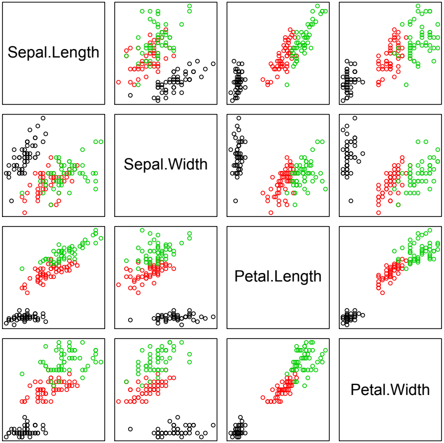
$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_k)$$

Allman, Matias, & Rhodes (2009) use a theorem by Kruskal (1976) to show that if:

- $f_{1k}, \ldots, f_{mk}$ are linearly independent for each k;
- $r \geq 3$

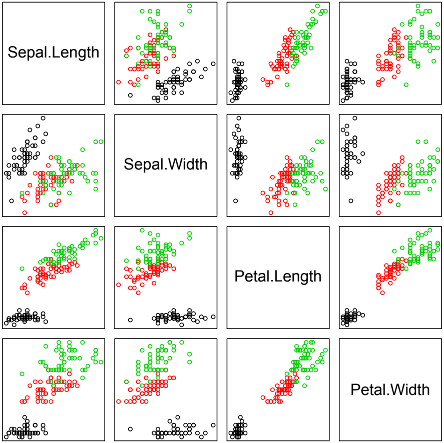
...then $g(\mathbf{x})$ uniquely determines all the λ_j and f_{jk} (up to label-switching).

Iris Data



- Well-known dataset for testing clustering methods. . .

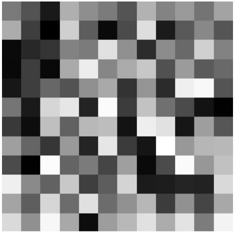
Iris Data



- Well-known dataset for testing clustering methods. . .
- . . . however, clearly conditional independence is violated for these data.

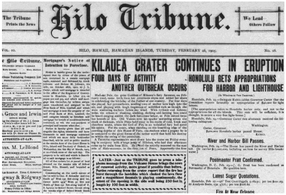
Larger Example: Classifying Image Excerpts

- **Goal:** Extracting intrinsic structure in images by clustering and finding complete set of efficient linear basis functions.



painting?

newspaper?



The Tribune Prints the News

Hilo Tribune.

We Lead Others Follow

KILAUEA CRATER CONTINUES IN ERUPTION

FOUR DAYS OF ACTIVITY

ANOTHER OUTBREAK OCCURS

HONOLULU GETS APPROPRIATIONS FOR HARBOR AND LIGHTHOUSE

River and Harbor Bill Passes.

Postmaster Pratt Confirmed.

Latest Sugar Quotations.

Outline

In some cases, conditional independence is too restrictive

Idea: Suppose $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ satisfy conditional independence but we observe $\mathbf{X}_i = A_j \mathbf{Y}_i$ when $\mathbf{Y}_i$ is in the jth mixture component.

- Here, $A_1, \ldots, A_m$ are $r \times r$ matrices.

Generalizing conditional independence

Observe $\mathbf{X}_i = A_j \mathbf{Y}_i$ when $\mathbf{Y}_i$ is in the jth mixture component:

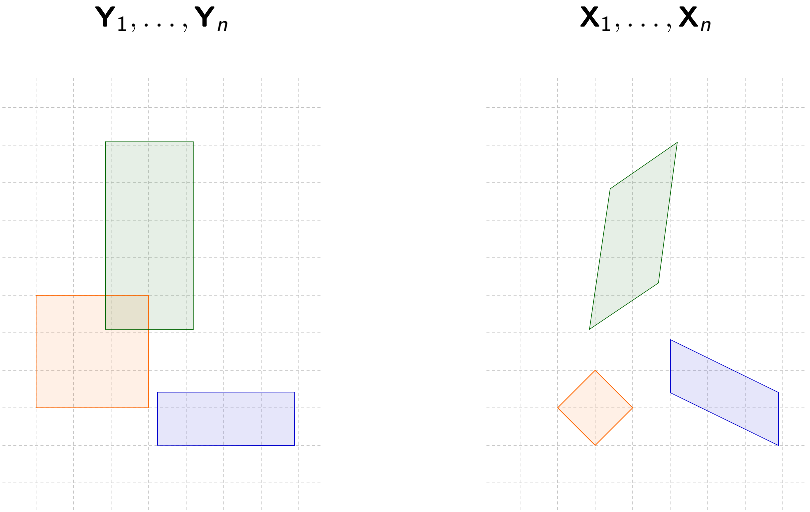
- $$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}),$$
- $$f_j(\mathbf{x}) = q_j(A_j^{-1}\mathbf{x})|\det A_j|^{-1}, \tag{2}$$
- $$q_j(\mathbf{y}) = \prod_{k=1}^{r} q_{jk}(y_k) \tag{3}$$

For identifiability, we assume

$$E(\mathbf{Y}_i \mathbf{Y}_i^\top) = I \tag{4}$$

Equations (2), (3), and (4) are common in the ICA (Independent Components Analysis) literature.

Component-wise ICA idea illustrated:



[Figure from Xiaotian Zhu's dissertation defense.]

We may eliminate the λ_j parameters

- Notation: Define f_{A} as

$$f_{\mathsf{A}}(\mathbf{x}) = f(\mathsf{A}^{-1}\mathbf{x})|\det \mathsf{A}|^{-1},$$

  for describing the density function of a linearly transformed random vector.

- Then we may write

$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}) = \sum_{j=1}^{m} (e_j)_{A_j}(\mathbf{x}).$$

Importantly, $e_j(\cdot) = \lambda_j q_j(\cdot)$ is not constrained to integrate to 1 as $f_j(\cdot)$ and $q_j(\cdot)$ are.

Estimation uses smoothed minimum penalized K-L divergence

Following Levine et al (2011) and Zhu and Hunter (2016), we propose estimating $\mathbf{e} = (e_1, e_2, ..., e_m)$ and $\mathrm{A} = (\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_m)$ by minimizing

$$\ell(\mathbf{e}, \mathrm{A}) =$$

$$\int g(\mathbf{x}) \log \left[g(\mathbf{x}) / \sum_{j=1}^{m} (\mathcal{N}_h e_j)_{\mathrm{A}_j}(\mathbf{x}) \right] d\mathbf{x} + \int \left[\sum_{j=1}^{m} (e_j)_{\mathrm{A}_j}(\mathbf{x}) \right] d\mathbf{x},$$

where $\mathcal{N}$ is the nonlinear smoother of Eggermont and LaRiccia (1995):

$$[\mathcal{N} f_j](\mathbf{x}) = \exp \int \frac{1}{h^r} K_r \left(\frac{\mathbf{x} - \mathbf{u}}{h^r} \right) \log f_j(\mathbf{u})\, d\mathbf{u}.$$

Finite sample version

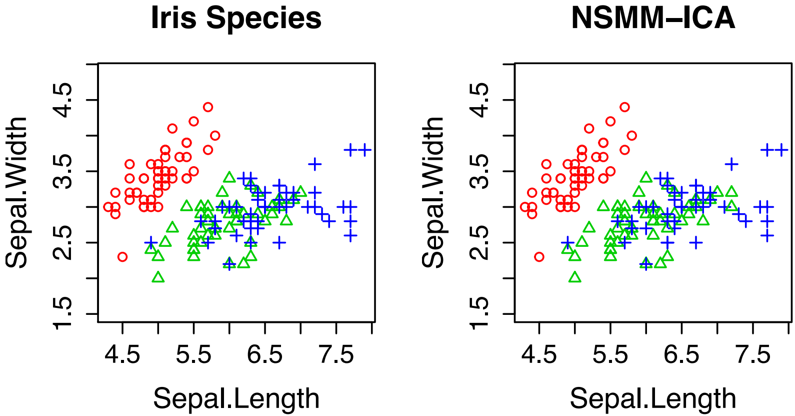
- The finite-sample version of the smoothed penalized K-L divergence is

$$\ell(\mathbf{e}, \mathrm{A}) = -\frac{1}{n}\sum_{i=1}^{n}\log\sum_{j=1}^{m}(\mathcal{N}_h e_j)_{\mathrm{A}_j}(\mathbf{x_i}) + \int\left[\sum_{j=1}^{m}(e_j)_{\mathrm{A}_j}(\mathbf{u})\right]\mathrm{d}\mathbf{u}.$$

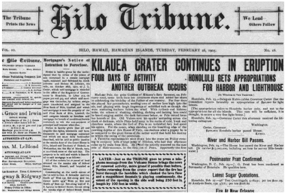
- Minimization is achieved via an MM algorithm.
- Theorem (Zhu and Hunter, 2016): If $\mathbf{e}$ is a solution to the main optimization problem, then

$$\int\sum_{j=1}^{m} e_j(\mathbf{u})\mathrm{d}\mathbf{u} = 1.$$

Iris flower data: NSMM-ICA misclassifies 7 out of 150.



Classifying Image excerpts



The Tribune Prints the News

Hilo Tribune.

We Lead Others Follow

HILO, HAWAII, HAWAIIAN ISLANDS, TUESDAY, FEBRUARY 28, 1905.

KILAUEA CRATER CONTINUES IN ERUPTION

FOUR DAYS OF ACTIVITY

ANOTHER OUTBREAK OCCURS

HONOLULU GETS APPROPRIATIONS FOR HARBOR AND LIGHTHOUSE

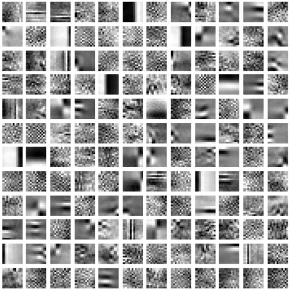
River and Harbor Bill Passes.

Postmaster Pratt Confirmed.

Latest Sugar Quotations.

- Data dimension: 10000×144
- Classification error rate: 1.2%

Basis functions for each mixture component visualized:



Summary

- Non- and semi-parametric mixture models are viable as long as parameters (including component densities) can be identified.
- Multiple results on identifiability exist, as do EM-like algorithms using kernel density estimation.

In the multivariate case:

- Algorithms and identifiability results have been derived for the conditional independence model.
- Conditional independence is too restrictive for some applications.
- Our idea: Alternate iterations of the NP-EM estimation algorithm with iterations of ICA.
- Currently implemented in `icamix` package for R.

References

- Allman, E. S., Matias, C., and Rhodes, J. A. (2009), Identifiability of parameters in latent structure models with many observed variables, *Annals of Statistics*, **37**: 3099–3132.
- Benaglia, T., Chauveau, D., and Hunter, D. R. (2009), An EM-like algorithm for semi- and non-parametric estimation in mutlivariate mixtures, *Journal of Computational and Graphical Statistics*, **18**: 505–526.
- Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009), mixtools: An R Package for Analyzing Finite Mixture Models, *Journal of Statistical Software*, **32**(6).
- Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006), Semiparametric estimation of a two-component mixture model, *Annals of Statistics*, **34**, 1204–1232.
- Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007), An EM algorithm for a semiparametric mixture model, *Computational Statistics and Data Analysis*, **51**: 5429–5443.
- Eggermont, P. P. B. and LaRiccia, V. N. (1995), Maximum Smoothed Density Estimation for Inverse Problems, *Annals of Statistics*, **23**, 199–220.
- Hall, P. and Zhou, X. H. (2003) Nonparametric estimation of component distributions in a multivariate mixture, *Annals of Statistics*, **31**: 201–224.
- Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007), Inference for mixtures of symmetric distributions, *Annals of Statistics*, **35**: 224–251.
- Hunter, D. R. and Young, D. S. (2012), "Semiparametric Mixtures of Regressions," *Journal of Nonparametric Statistics*, **24** (1): 19–38.
- Levine, M., Hunter, D. R., and Chauveau, D. (2011), Maximum Smoothed Likelihood for Multivariate Mixtures, *Biometrika*, **98** (2): 403–416.
- Kruskal, J. B. (1976), More Factors Than Subjects, Tests and Treatments: An Indeterminacy Theorem for Canonical Decomposition and Individual Differences Scaling, *Psychometrika*, **41**: 281–293.
- Qin, J. and Leung, D. H.-Y. (2006), Semiparametric analysis in conditionally independent multivariate mixture models, unpublished manuscript.
- Zhu, X. and Hunter, D. R. (2016), Theoretical Grounding for Estimation in Conditional Independence Multivariate Finite Mixture Models, *Journal of Nonparametric Statistics*, **28**: 683–701.
- Zhu, X. and Hunter, D. R. (2015), icamix: Estimation of ICA Mixture Models, R package version 1.0.2.