Why is Mixture Modelling so popular?

Tony Robinson

Department of Mathematical Sciences
University of Bath

15th May, 2008

Outline

Outline

1. Heterogeneity - one model won't do!

Outline

1. Heterogeneity - one model won't do!
2. Basic mixture formulation.

Outline

1. Heterogeneity - one model won't do!
2. Basic mixture formulation.
3. Ways to fit.

Outline

1. Heterogeneity - one model won't do!
2. Basic mixture formulation.
3. Ways to fit.
4. Inferential difficulties.

Outline

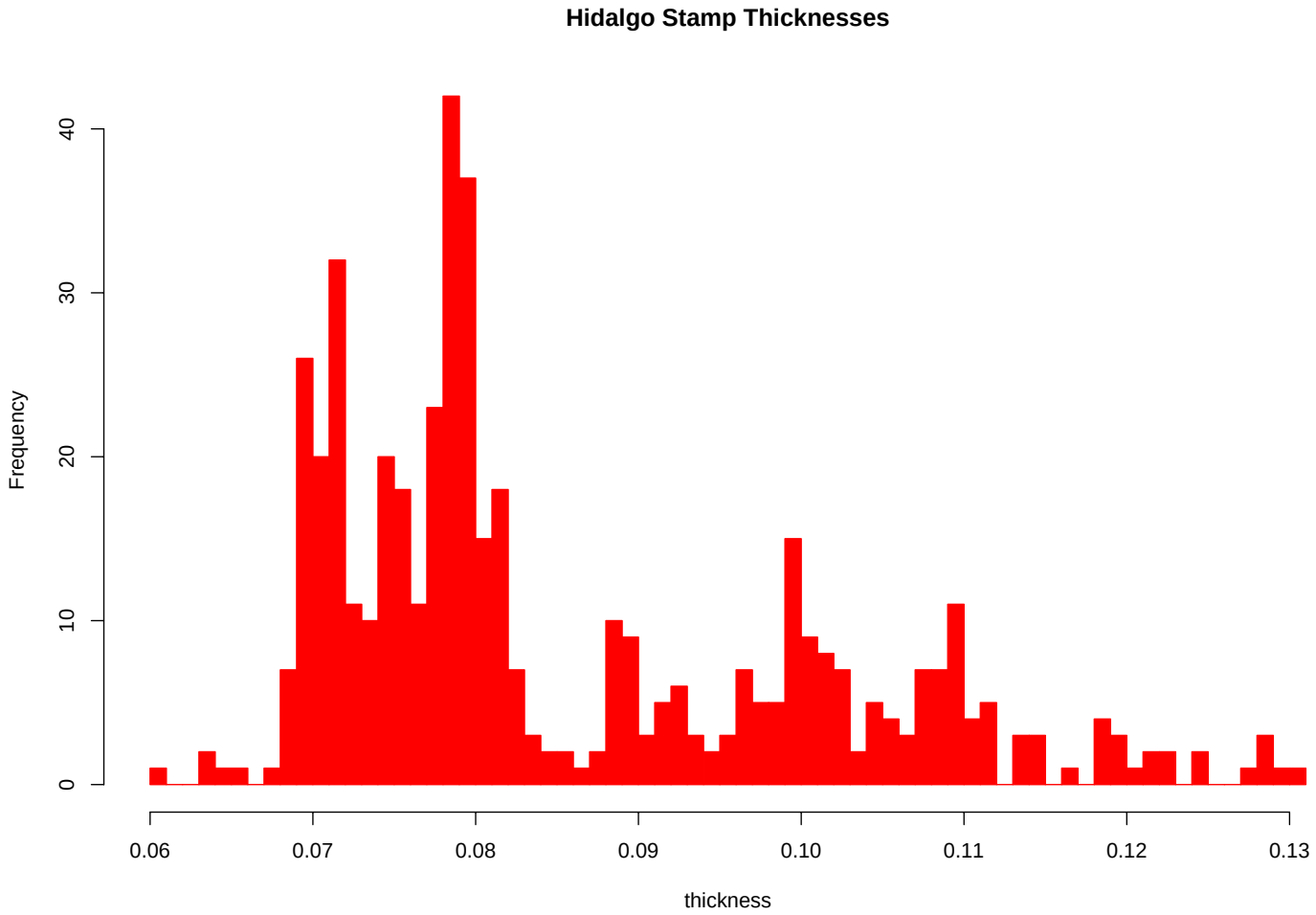
1. Heterogeneity - one model won't do!
2. Basic mixture formulation.
3. Ways to fit.
4. Inferential difficulties.
5. Cautionary example.

Introduction - Model Heterogeneity.

Much of the data we encounter does not seem to be appropriately fitted by those nice simple probability models we learnt about “in school”.

Introduction - Model Heterogeneity.

Much of the data we encounter does not seem to be appropriately fitted by those nice simple probability models we learnt about “in school”.



Finite Mixture Models

Finite Mixture Models

Idea is to model complex data as a finite mixture of component models, often of same type.

Finite Mixture Models

Idea is to model complex data as a finite mixture of component models, often of same type.
Loosely, model M is a weighted mixture of component models $\{M_i\}$, if

$$M = \sum_{i=1}^{g} w_i M_i$$

w_i represents the proportion of the data "explained" by M_i.

Finite Mixture Models

Idea is to model complex data as a finite mixture of component models, often of same type.
Loosely, model M is a weighted mixture of component models $\{M_i\}$, if

$$M = \sum_{i=1}^{g} w_i M_i$$

w_i represents the proportion of the data "explained" by M_i.
Example : Density estimation

$$f(\mathbf{x}_j) = \sum_{i=1}^{g} w_i f_i(\mathbf{x}_j),$$

$f_i(\mathbf{x}_j)$ are "standard" densities and $0 \leq w_i \leq 1,$ and $\sum_{i=1}^{g} w_i = 1.$

For the Hidalgo stamp thickness data, one simple objective might be density estimation.

For the Hidalgo stamp thickness data, one simple objective might be density estimation.
We could model the thicknesses as a mixture of Gaussian distributions.

$$f(\mathbf{x}_j) = \sum_{i=1}^{g} w_i \phi_i(\mathbf{x}_j, \mu_i, \sigma_i^2),$$

For the Hidalgo stamp thickness data, one simple objective might be density estimation.
We could model the thicknesses as a mixture of Gaussian distributions.

$$f(\mathbf{x}_j) = \sum_{i=1}^{g} w_i \phi_i(\mathbf{x}_j, \mu_i, \sigma_i^2),$$

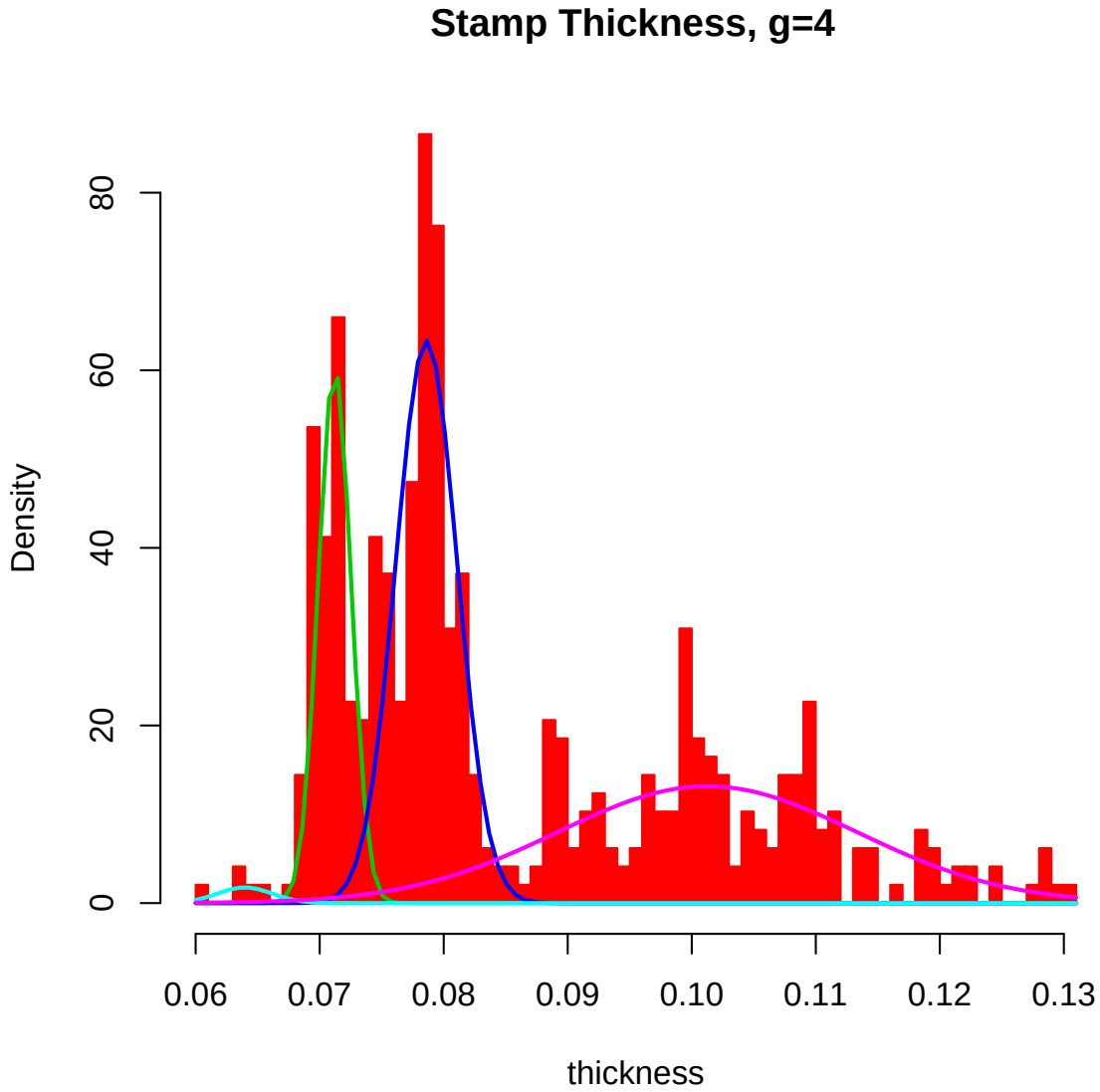
But this raises an immediate question?

For the Hidalgo stamp thickness data, one simple objective might be density estimation.
We could model the thicknesses as a mixture of Gaussian distributions.

$$f(\mathbf{x}_j) = \sum_{i=1}^{g} w_i \phi_i(\mathbf{x}_j, \mu_i, \sigma_i^2),$$

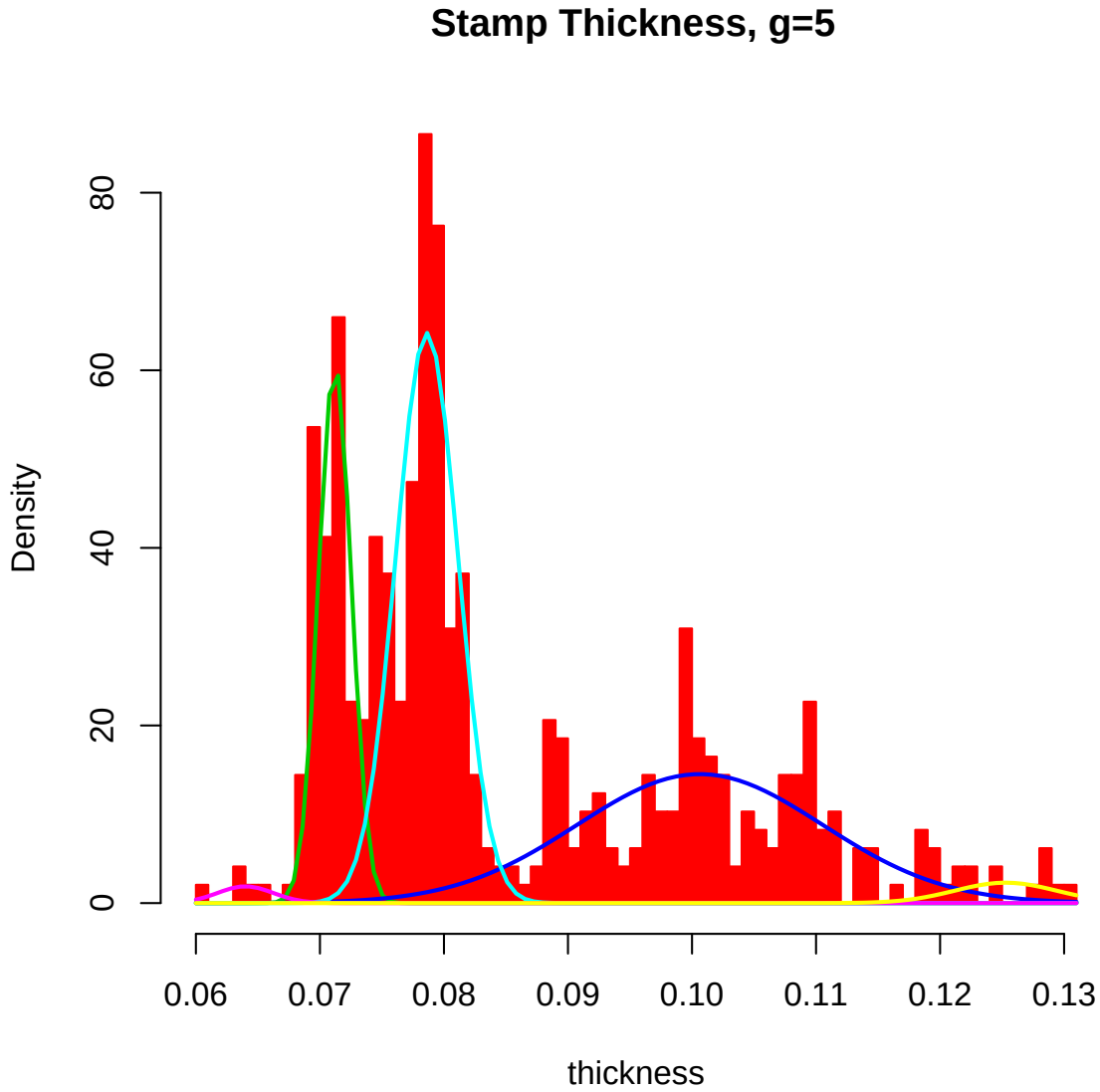
But this raises an immediate question?
Do we *know* how many component models there are in the mixture? Or, do we need to find out?

Four?


Stamp Thickness, g=4
Density
0
20
40
60
80
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
thickness

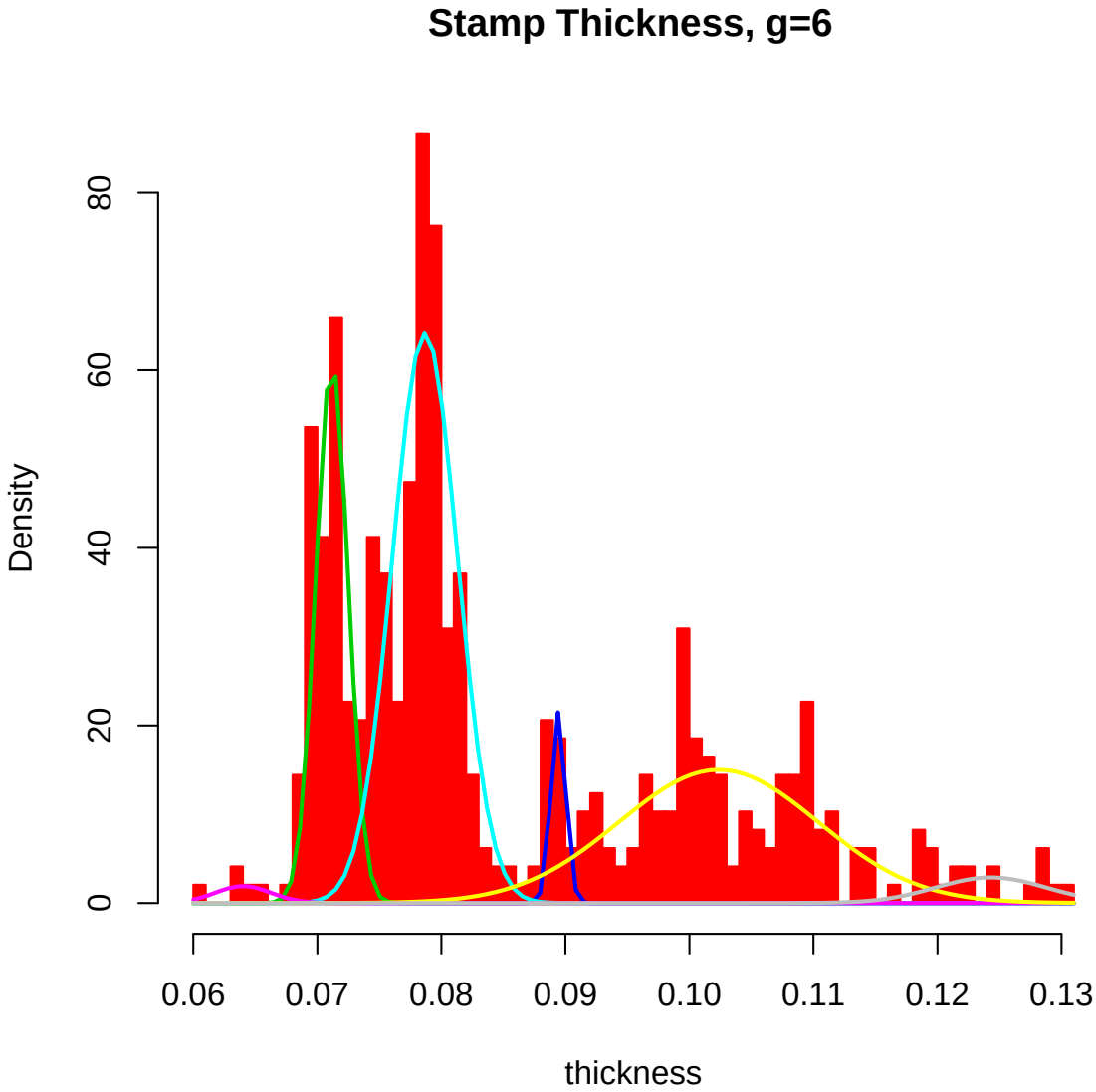
Four?

Five?


Stamp Thickness, g=5
Density
0
20
40
60
80
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
thickness

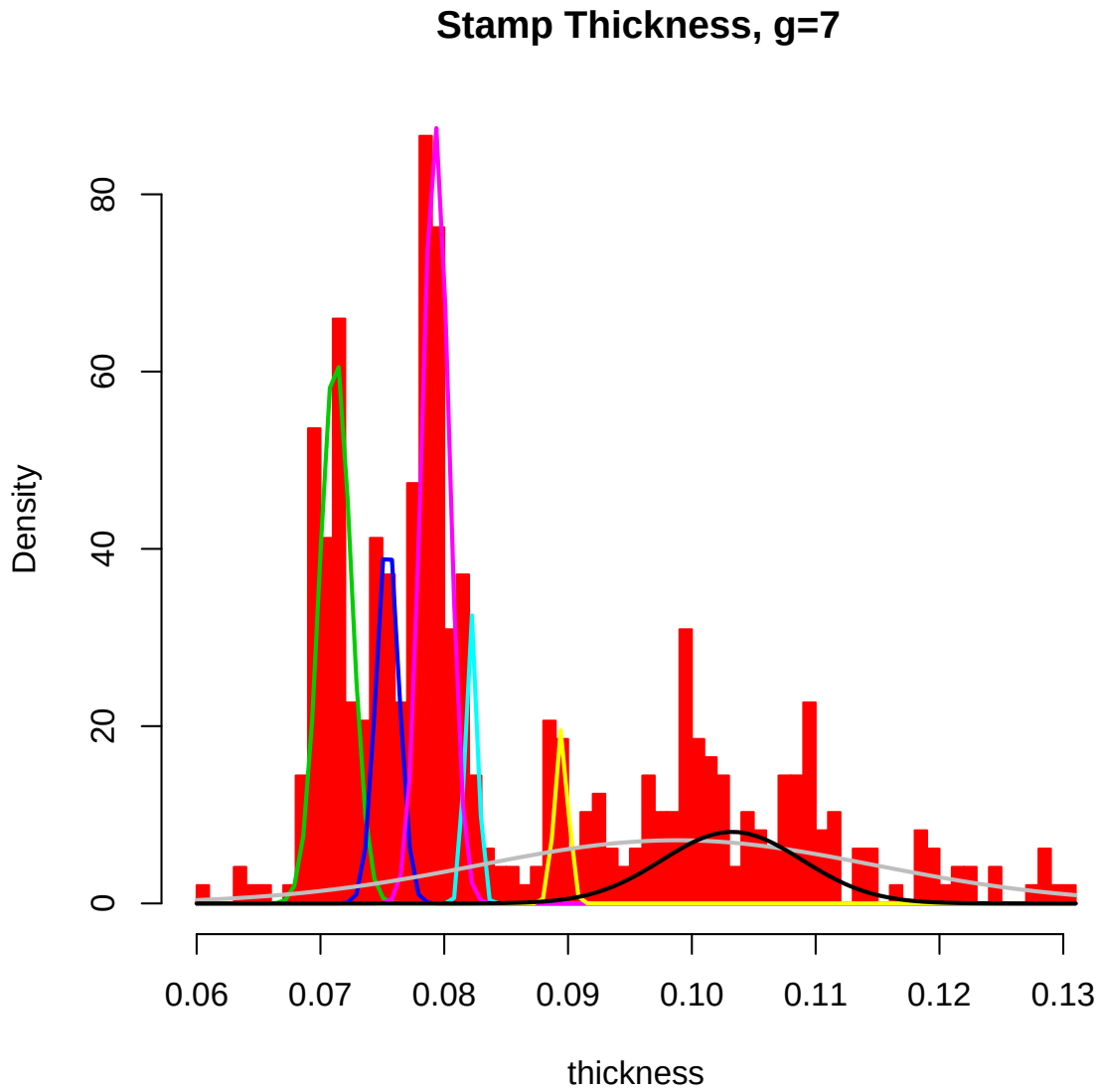
Five?

Six?


Stamp Thickness, g=6
Density
0
20
40
60
80
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
thickness

Six?

Seven?


Stamp Thickness, g=7
Density
0
20
40
60
80
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
thickness

Seven?

Model-Based Clustering

Model-Based Clustering

The stamp data paper is likely to come from several different origins.

Model-Based Clustering

The stamp data paper is likely to come from several different origins.
Mixture modelling provides a natural framework for this situation.

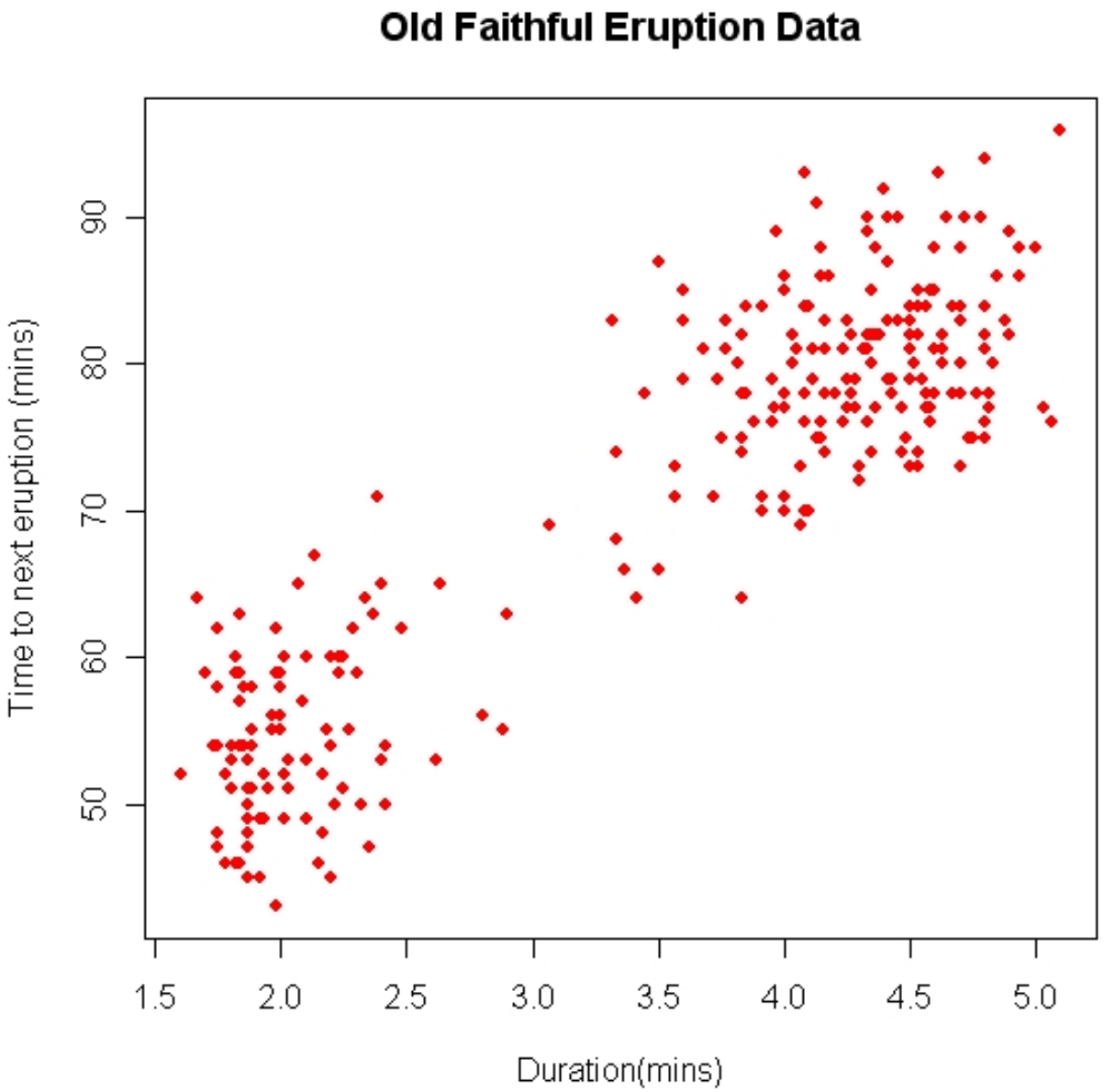
Model-Based Clustering

The stamp data paper is likely to come from several different origins.
Mixture modelling provides a natural framework for this situation.
At its simplest, each component in the mixture represents a cluster within the total population.

Model-Based Clustering

The stamp data paper is likely to come from several different origins.
Mixture modelling provides a natural framework for this situation.
At its simplest, each component in the mixture represents a cluster within the total population.
Assumes clusters are individually well fitted by the models $\{M_i\}$.

Old Faithful Data



More Applications

Mixtures of regressions, linear Models, GLMs, survival,...

EG. Hurn *et al.* (2000)

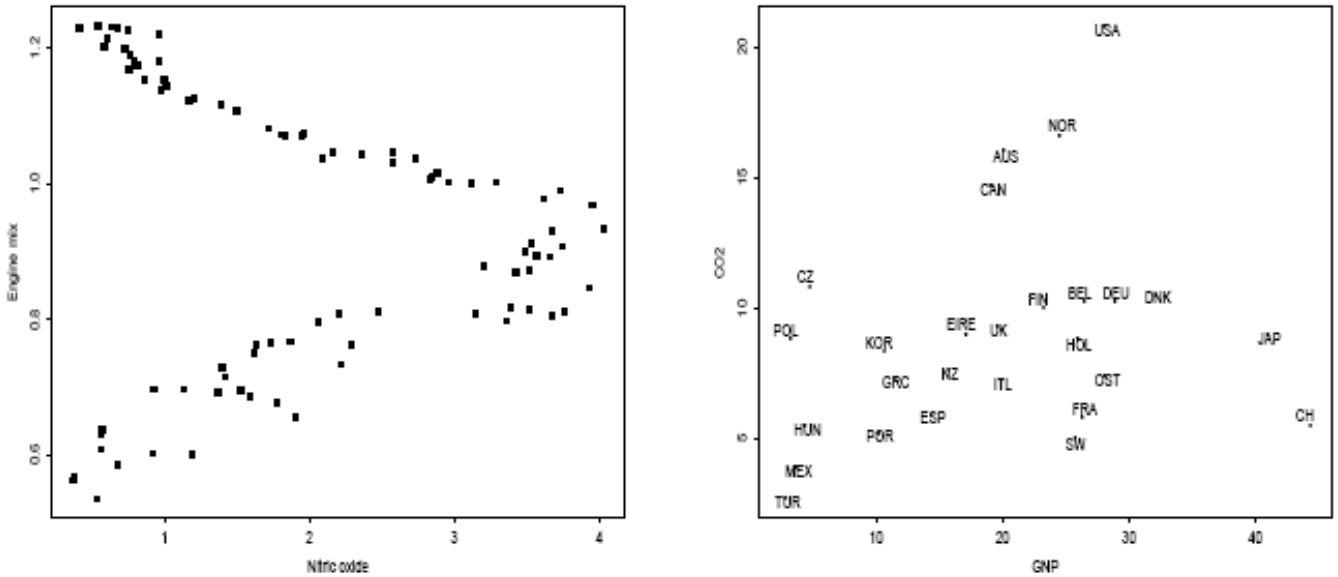
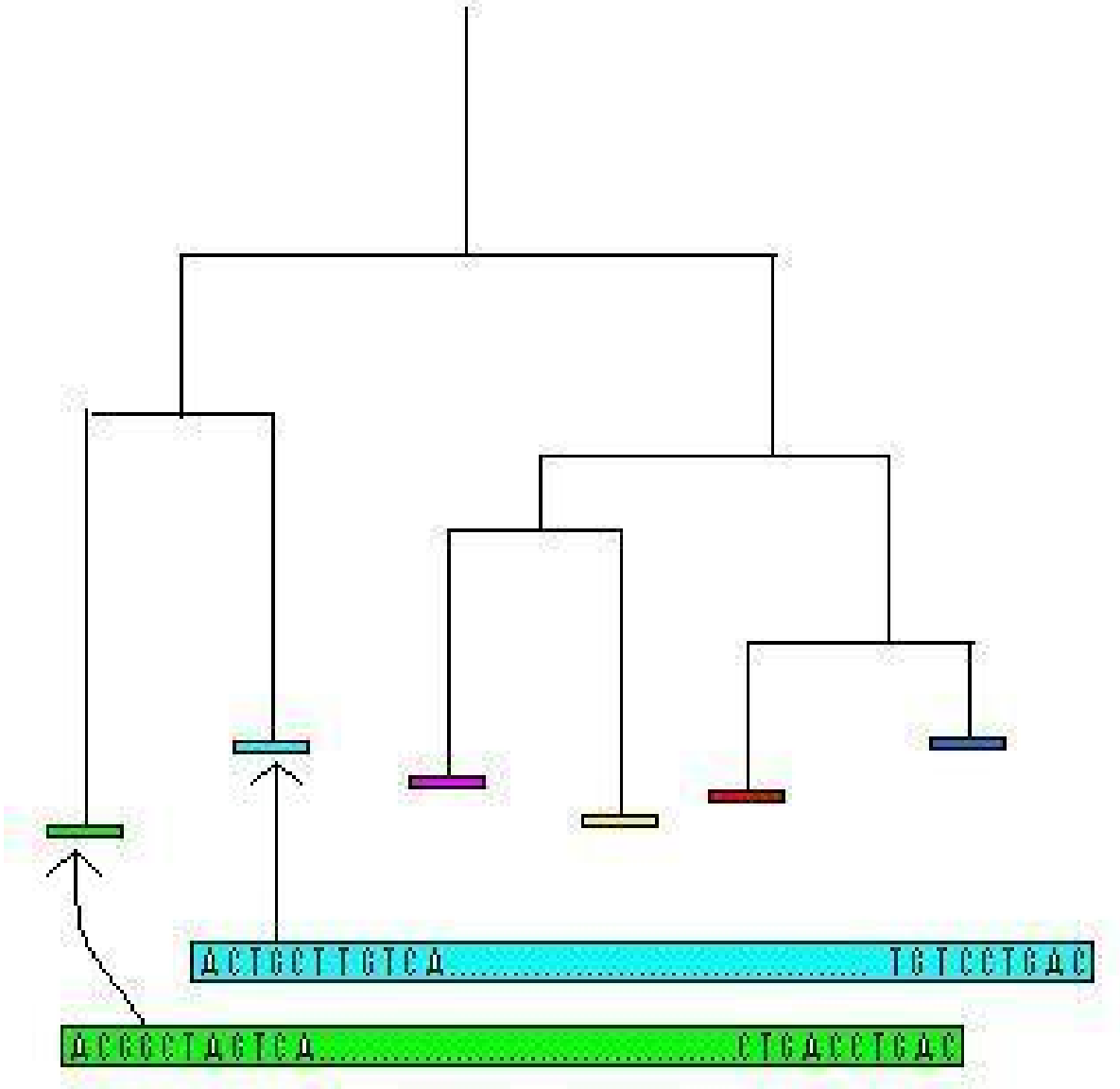


Fig. 2. *(left)* Equivalence ratio against exhaust nitric oxide concentration (*Source:* Hurvich *et al.*, 1998); *(right)* representation of the GNP and CO2 emission levels in 1996 for various countries (*Source:* OECD).

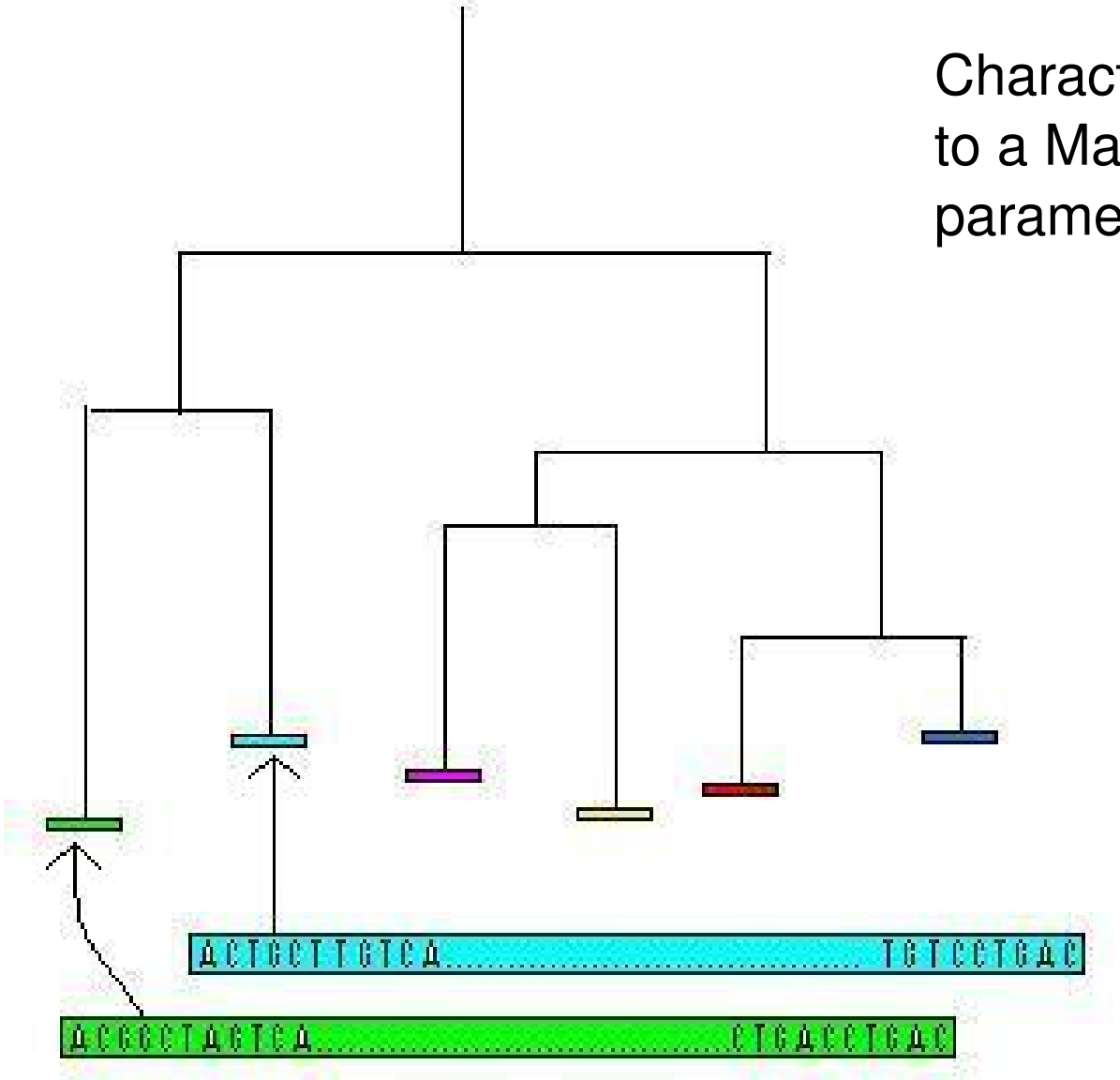
More Applications

Mixtures of Phylogenetic Models (Evolutionary tree of species)



More Applications

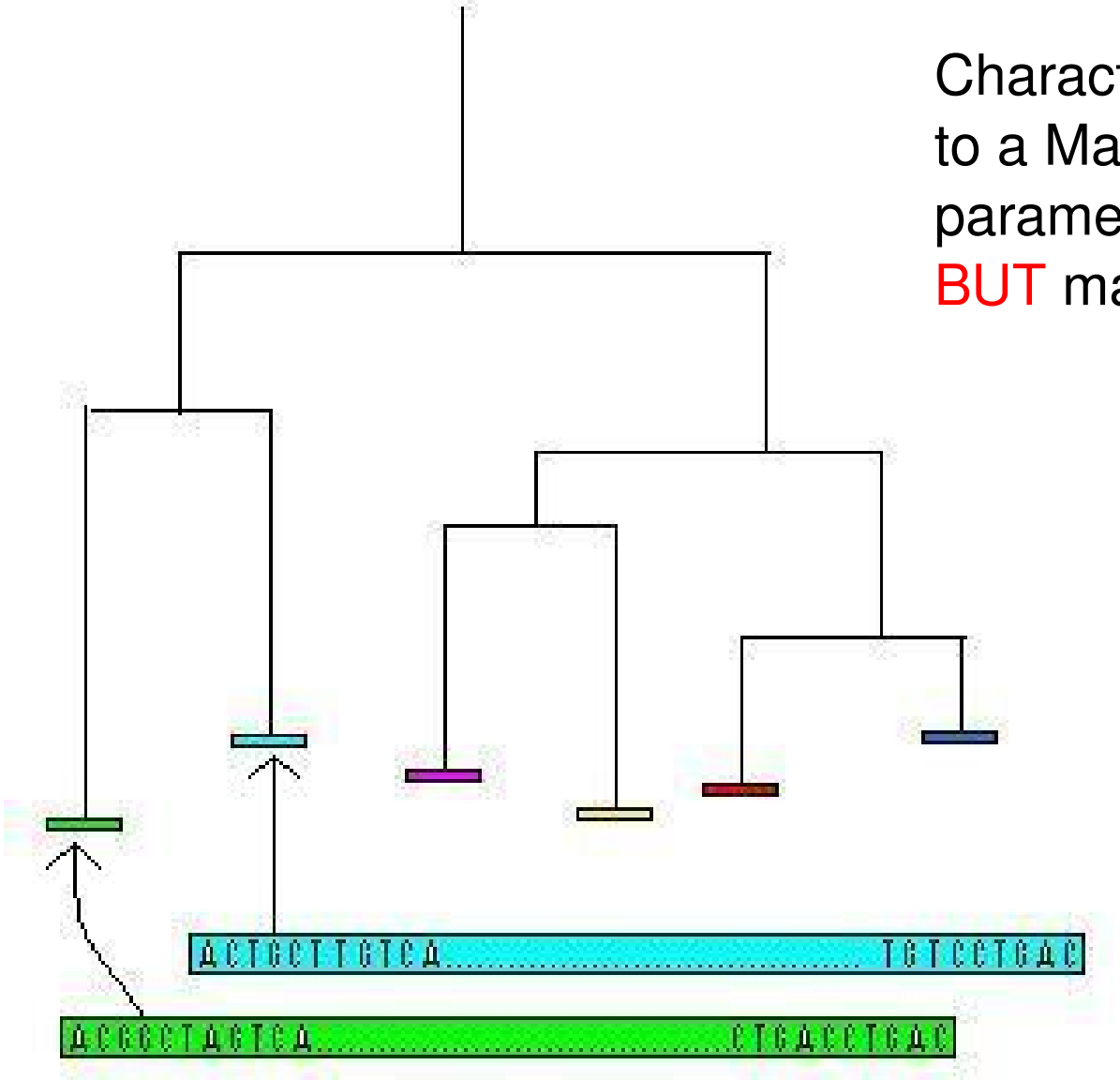
Mixtures of Phylogenetic Models (Evolutionary tree of species)



Characters (ACGT) evolve according to a Markov process Θ involving parameters such as mutation rates.

More Applications

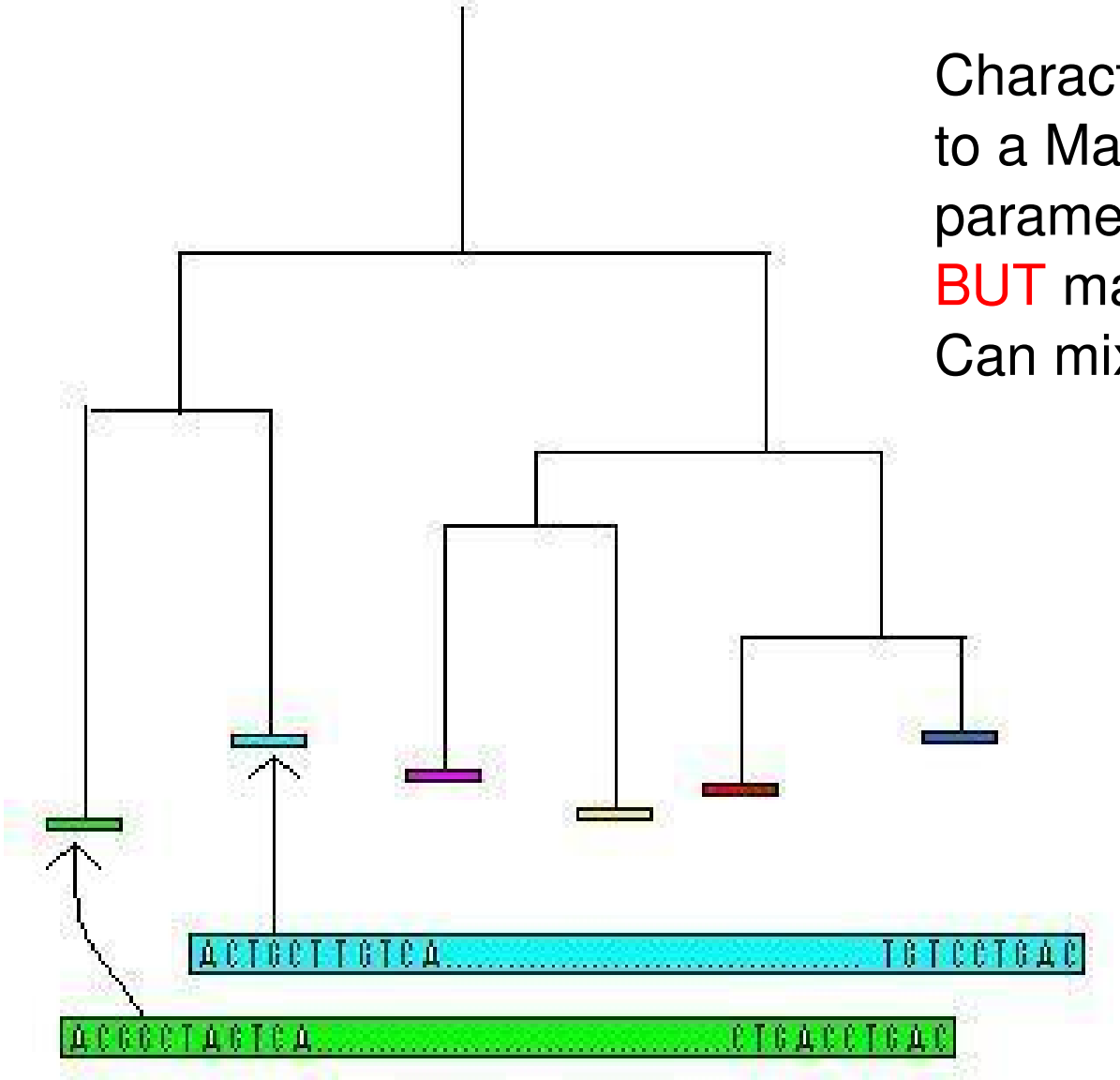
Mixtures of Phylogenetic Models (Evolutionary tree of species)



Characters (ACGT) evolve according to a Markov process Θ involving parameters such as mutation rates.
BUT many sources of heterogeneity

More Applications

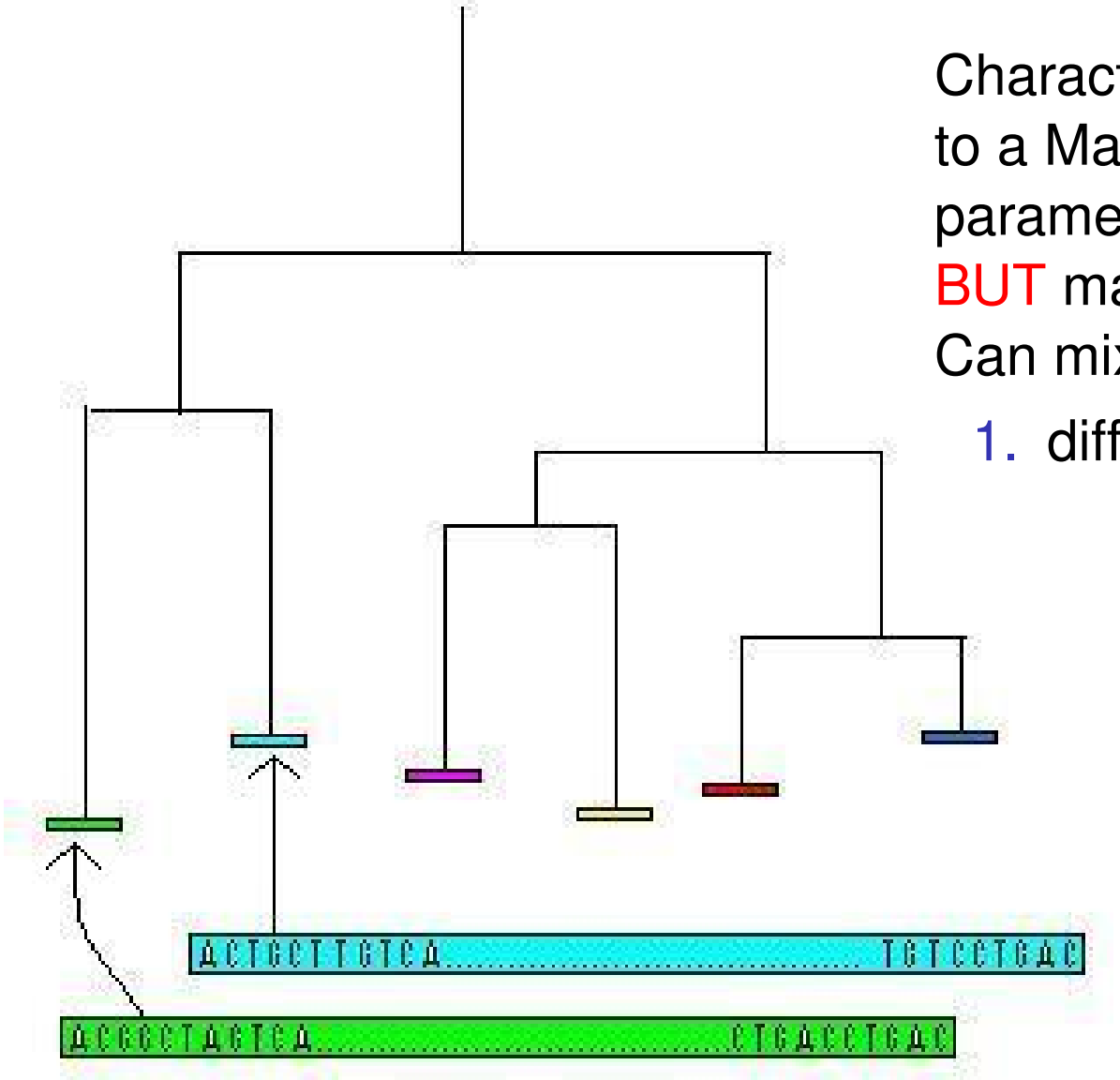
Mixtures of Phylogenetic Models (Evolutionary tree of species)



Characters (ACGT) evolve according to a Markov process Θ involving parameters such as mutation rates.
BUT many sources of heterogeneity
Can mix any or all of

More Applications

Mixtures of Phylogenetic Models (Evolutionary tree of species)

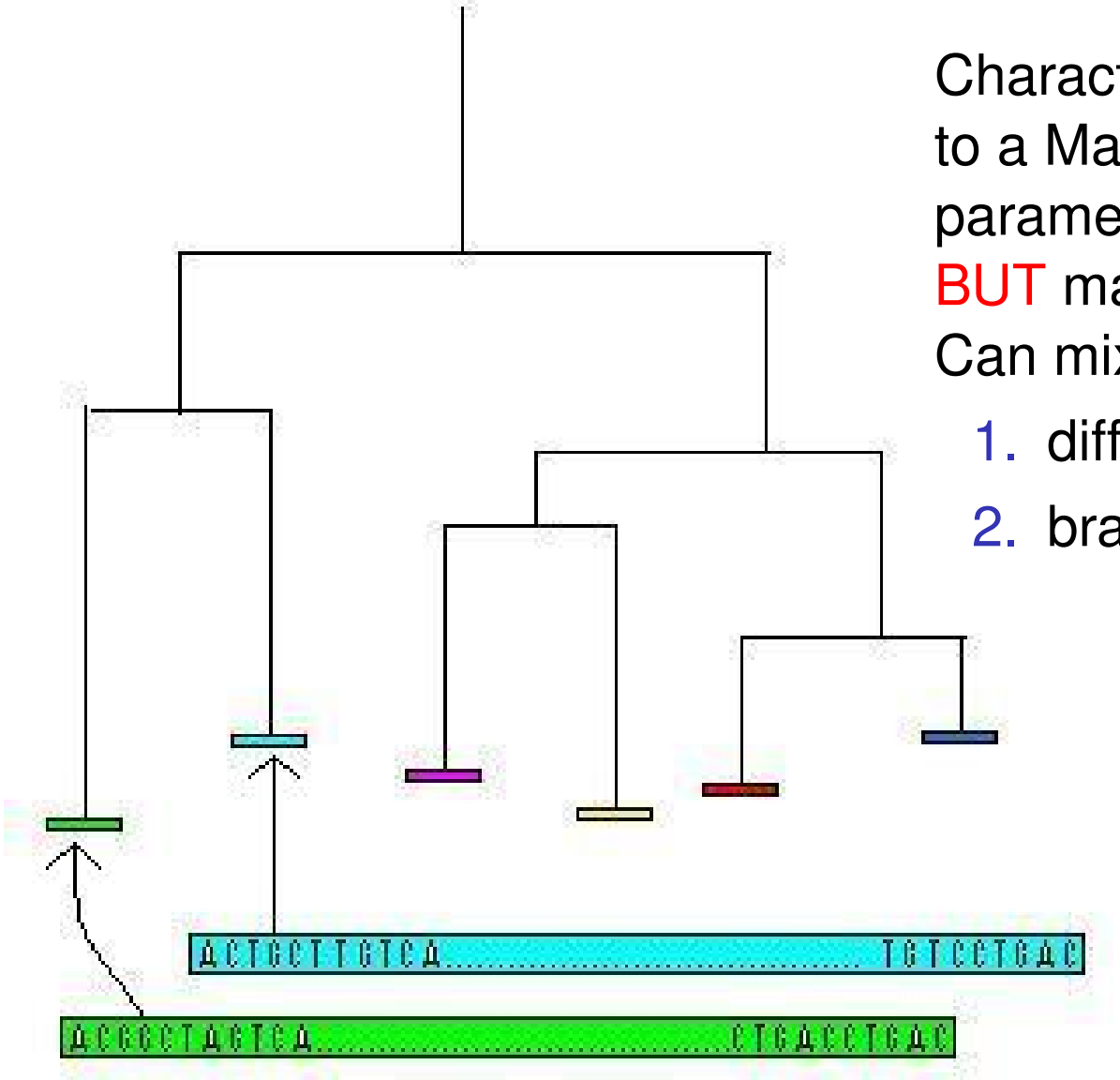


Characters (ACGT) evolve according to a Markov process Θ involving parameters such as mutation rates.
BUT many sources of heterogeneity
Can mix any or all of

1. different Θs

More Applications

Mixtures of Phylogenetic Models (Evolutionary tree of species)

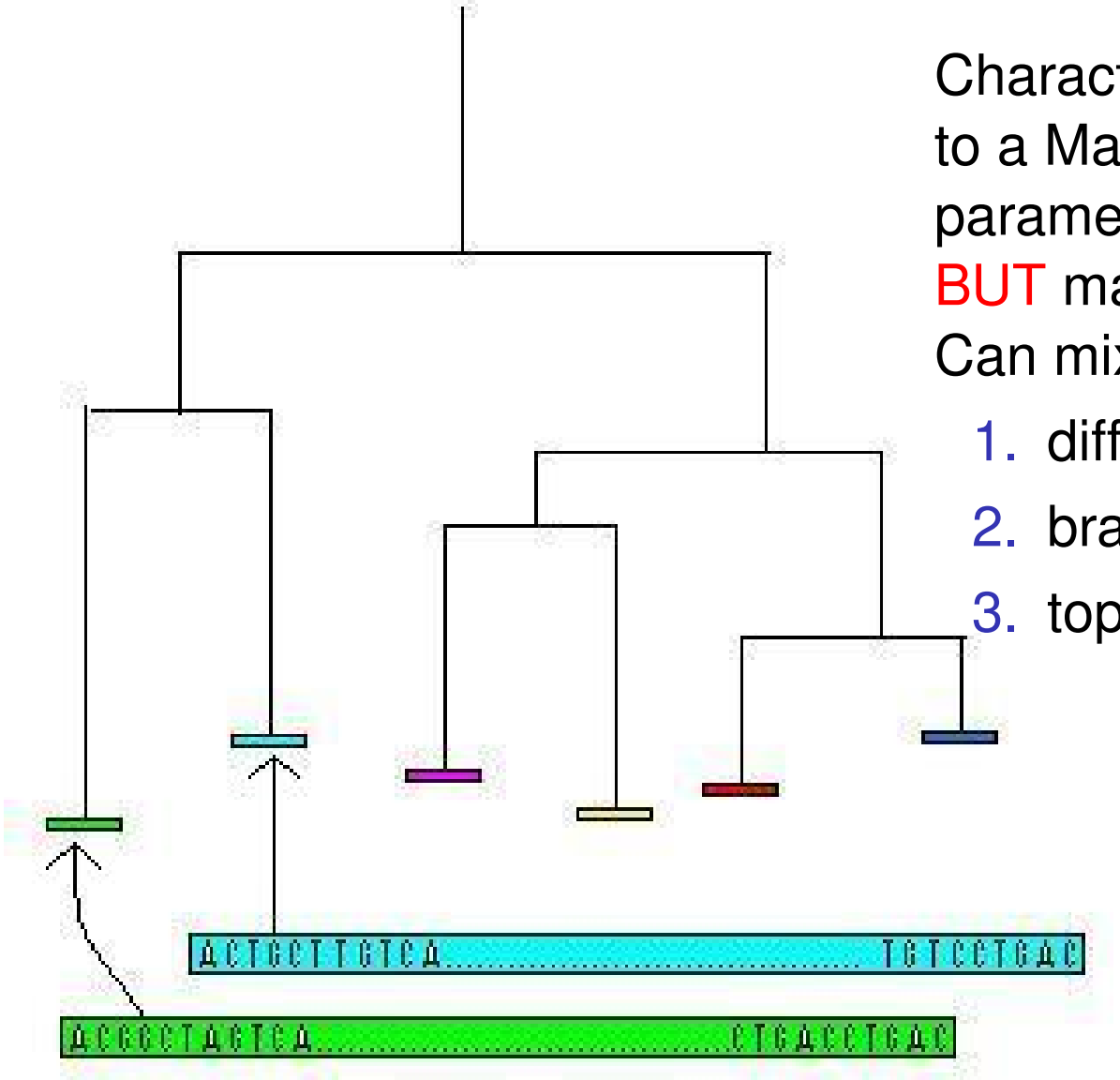


Characters (ACGT) evolve according to a Markov process Θ involving parameters such as mutation rates.
BUT many sources of heterogeneity
Can mix any or all of

1. different Θs
2. branch lengths

More Applications

Mixtures of Phylogenetic Models (Evolutionary tree of species)



Characters (ACGT) evolve according to a Markov process Θ involving parameters such as mutation rates.
BUT many sources of heterogeneity
Can mix any or all of

1. different Θs
2. branch lengths
3. topologies

More Applications

Social Networks Handcock *et al*(2006)

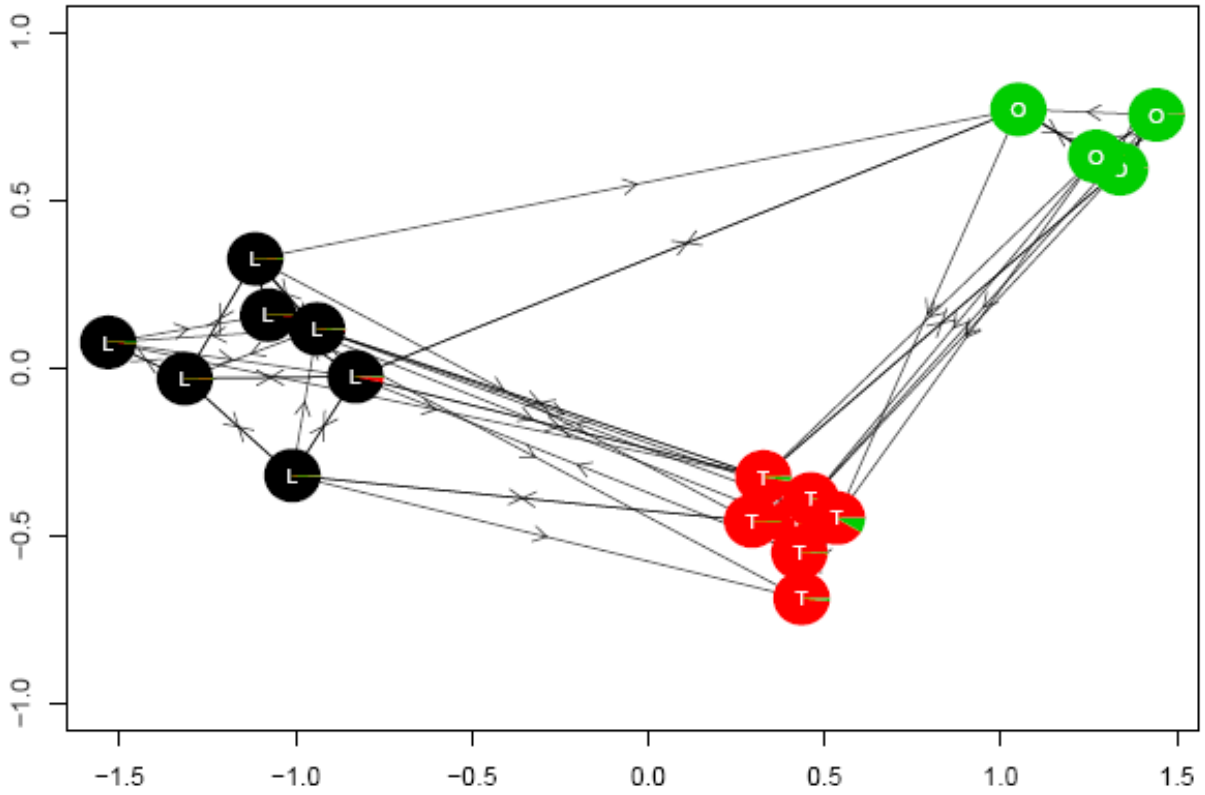


Figure 3: Estimates of clusters and latent positions for the relationship between monks within a monastery from the Bayesian estimation of the LPCM. The probability of assignment to each latent clusters is shown by a colored pie chart.

Fitting Mixture Models

Two popular methods of fitting.

Fitting Mixture Models

Two popular methods of fitting.

1. EM Algorithm

Fitting Mixture Models

Two popular methods of fitting.

1. EM Algorithm
2. Markov chain Monte Carlo.

Fitting Mixture Models

Two popular methods of fitting.

1. EM Algorithm
2. Markov chain Monte Carlo.

The EM algorithm involves assessing to which component j of the mixture each datum is expected to belong. Once this is established for all the data, fitting by usual MLE proceeds for the parameters of model j.
McMC may also utilise this device for completing the "missing" part of the data.

Allocation Variables

- **Latent allocation variables** describe which observations are assigned to each of the current components at each iteration of the EM algorithm or McMC sampler.
  In the case of McMC these provide a sample cluster configuration per McMC iteration.

Allocation Variables

- **Latent allocation variables** describe which observations are assigned to each of the current components at each iteration of the EM algorithm or McMC sampler.
  In the case of McMC these provide a sample cluster configuration per McMC iteration.
- Set $Z_j = i$ if observation j is allocated to component i

Allocation Variables

- **Latent allocation variables** describe which observations are assigned to each of the current components at each iteration of the EM algorithm or McMC sampler.
  In the case of McMC these provide a sample cluster configuration per McMC iteration.
- Set $Z_j = i$ if observation j is allocated to component i
- Resulting allocation vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$

Any Problems?

Model selection, inference and interpretation can be complicated by several factors.

Any Problems?

Model selection, inference and interpretation can be complicated by several factors.

- What is g?
- g large $\Rightarrow$ many parameters.
- Identifiability?

Label Switching

Label Switching

- In a mixture model, if all the g components belong to the same parametric family, then the mixture density is invariant under the $g!$ permutations of the component labels.

Label Switching

- In a mixture model, if all the g components belong to the same parametric family, then the mixture density is invariant under the $g!$ permutations of the component labels.
- The likelihood

$$Lik = \prod_{i=1}^{n} \{w_1 f(\mathbf{x}_i; \theta_1) + \ldots + w_g f(\mathbf{x}_i; \theta_g)\}$$

is the same for all permutations of the labels $(1, 2, 3, ..., g)$.

Label Switching

- In a mixture model, if all the g components belong to the same parametric family, then the mixture density is invariant under the $g!$ permutations of the component labels.
- The likelihood

$$Lik = \prod_{i=1}^{n} \{w_1 f(\mathbf{x}_i; \theta_1) + \ldots + w_g f(\mathbf{x}_i; \theta_g)\}$$

  is the same for all permutations of the labels $(1, 2, 3, ..., g)$.
- The term *label switching* is used to describe the invariance of the likelihood under the relabelling of the components.

Label Switching

- In a mixture model, if all the g components belong to the same parametric family, then the mixture density is invariant under the $g!$ permutations of the component labels.
- The likelihood

$$Lik = \prod_{i=1}^{n} \{w_1 f(\mathbf{x}_i; \theta_1) + \ldots + w_g f(\mathbf{x}_i; \theta_g)\}$$

  is the same for all permutations of the labels $(1, 2, 3, ..., g)$.
- The term *label switching* is used to describe the invariance of the likelihood under the relabelling of the components.
- Often handled by imposing constraints, e.g label components in increasing order of weight w_j. Not always satisfactory.

Identifiability v Nonidentifiability

Identifiability v Nonidentifiability

- For example, consider the allocation vectors $(4,4,3,3,4,2,2,3,1,3)$ and $(2,2,1,1,2,3,3,1,4,1)$ these are different models if each component can be identified from some information.

Identifiability v Nonidentifiability

- For example, consider the allocation vectors $(4,4,3,3,4,2,2,3,1,3)$ and $(2,2,1,1,2,3,3,1,4,1)$ these are different models if each component can be identified from some information.
- However these two vectors define the same partition of the data and so are identical models from the clustering viewpoint. Therefore we would like some unique representation of them that identifies their common partition, i.e. $\{\{1,2,5\},\{3,4,8,10\},\{6,7\},\{9\}\}$.

RGF Representation

RGF Representation

- One simple solution is to relabel the allocations in increasing order as new components make an appearance, starting with label 0 or 1 for the first observation.

RGF Representation

- One simple solution is to relabel the allocations in increasing order as new components make an appearance, starting with label 0 or 1 for the first observation.
- The restricted growth function representation of the two allocation vectors above is

$$\{1,1,2,2,1,3,3,2,4,2\}$$

RGF Representation

- One simple solution is to relabel the allocations in increasing order as new components make an appearance, starting with label 0 or 1 for the first observation.
- The restricted growth function representation of the two allocation vectors above is

$$\{1,1,2,2,1,3,3,2,4,2\}$$

- Given allocation vectors from a sampler possibly suffering from label switching, we convert them to an unambiguous sample from space of possible partitions.

Example, Old Faithful Data

Example, Old Faithful Data

- Using MVN mixtures, we obtained a sample of 2000 allocations from a reversible jump McMC sampler with a burn-in of 300000 iterations and a subsequent sample of 100000 iterations thinned every 50.

Example, Old Faithful Data

- Using MVN mixtures, we obtained a sample of 2000 allocations from a reversible jump McMC sampler with a burn-in of 300000 iterations and a subsequent sample of 100000 iterations thinned every 50.
- So estimating g as well as Gaussian parameters and weights.

Example, Old Faithful Data

- Using MVN mixtures, we obtained a sample of 2000 allocations from a reversible jump McMC sampler with a burn-in of 300000 iterations and a subsequent sample of 100000 iterations thinned every 50.
- So estimating g as well as Gaussian parameters and weights.
- Standard approach to clustering is typically:
  1. Estimate likely number of components.
  2. Given g, estimate likely clustering.
  3. This conditional approach can easily mislead.

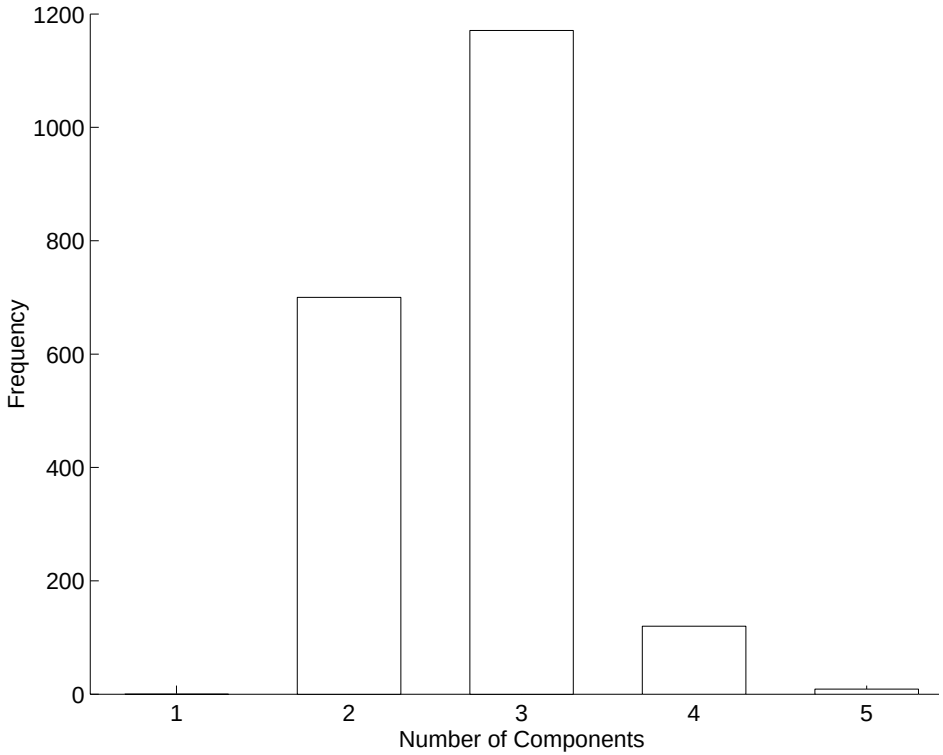
Old Faithful “Standard” Clustering

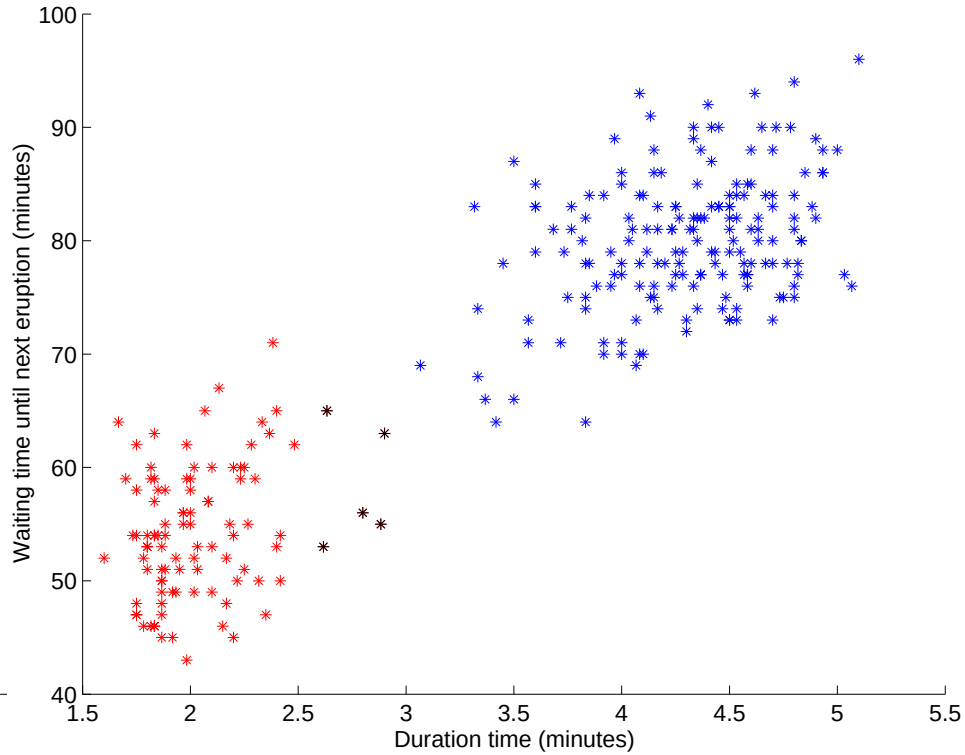
Old Faithful “Standard” Clustering

The figure below shows the posterior distribution of the number of components k which has a prominent mode at 3. Alongside is the classification into 3 clusters obtained by standard hierarchical clustering using the number of times each pair of observations appear in the same sample cluster as a distance measure. (O’Hagan)

Old Faithful "Standard" Clustering

The figure below shows the posterior distribution of the number of components k which has a prominent mode at 3. Alongside is the classification into 3 clusters obtained by standard hierarchical clustering using the number of times each pair of observations appear in the same sample cluster as a distance measure. (O'Hagan)





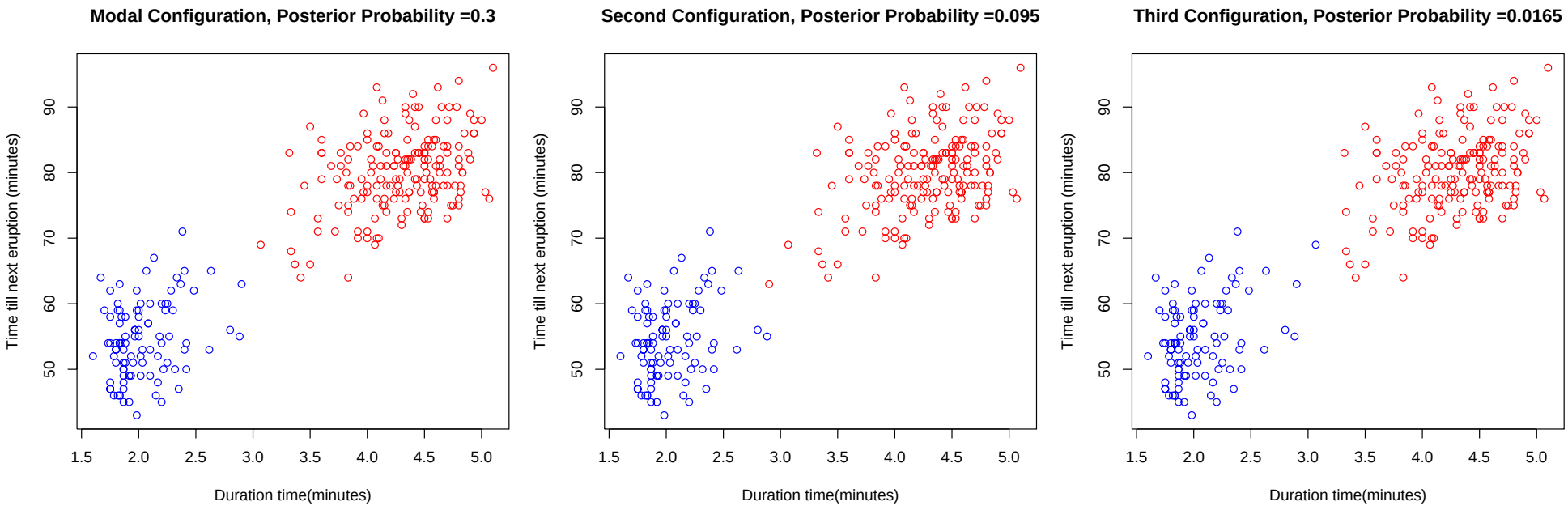
Old Faithful, Common Partitions

Old Faithful, Common Partitions

The figure below depicts the 3 most commonly occurring configurations - all contain just 2 components.

Old Faithful, Common Partitions

The figure below depicts the 3 most commonly occurring configurations - all contain just 2 components.



Can we do more?

- Simple to explore the commonly occuring partitions.
- But what about variation in the sampled partition values?
- These live in a very complex and high dimensional discrete space.
- Could we get some insight into nature of this sample distribution?
- Dissimilarity between each pair of partitions - number of pairs of observations that agree in the two partions, ie. either both in same group or both in different groups.

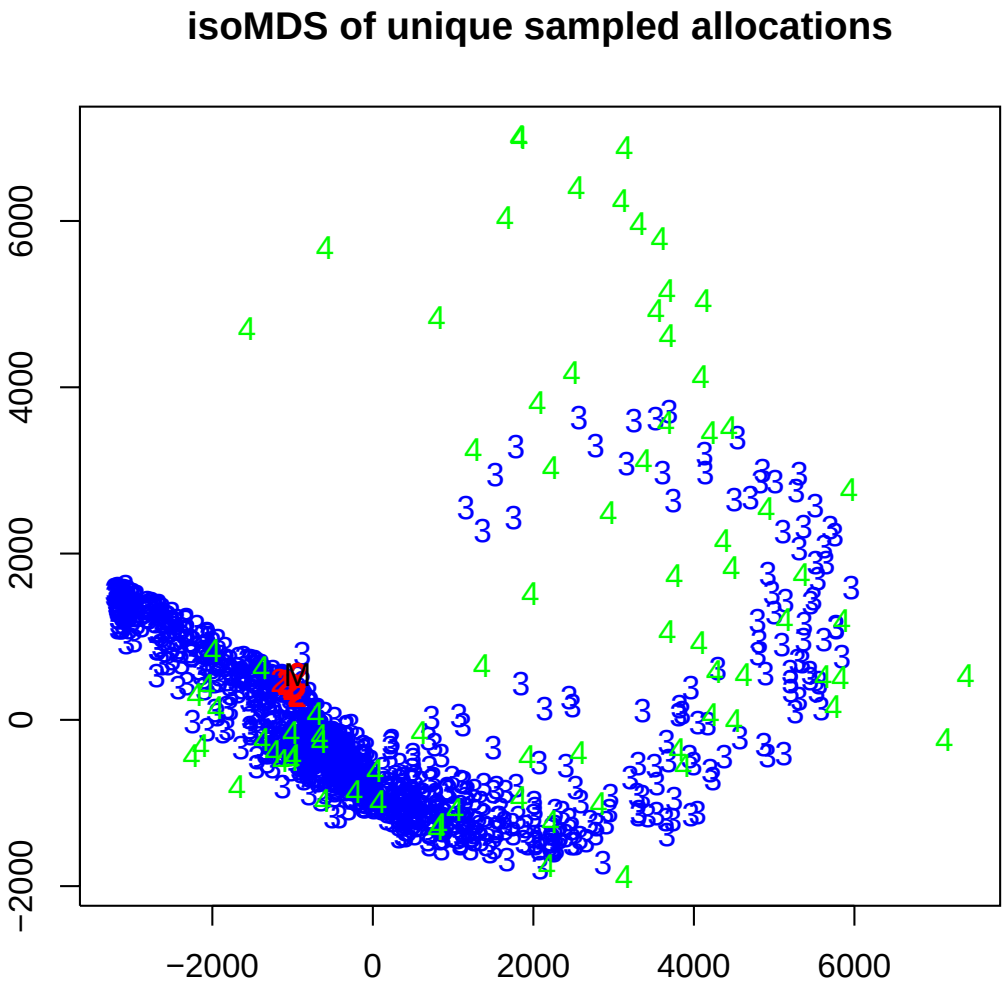
MDS plot of Old Faithful partitions

MDS plot of Old Faithful partitions

Here is a 2D configuration obtained from a nonmetric multidimensional scaling based on the computed dissimilarities between the 1172 sampled cluster configurations.

MDS plot of Old Faithful partitions

Here is a 2D configuration obtained from a nonmetric multidimensional scaling based on the computed dissimilarities between the 1172 sampled cluster configurations.



Summary

- Mixture models very attractive for model heterogeneous data.

Summary

- Mixture models very attractive for model heterogeneous data.
- Wide applicability.

Summary

- Mixture models very attractive for model heterogeneous data.
- Wide applicability.
- In most applications, it matters, to some extent, which M_i each observation belongs to.

Summary

- Mixture models very attractive for model heterogeneous data.
- Wide applicability.
- In most applications, it matters, to some extent, which M_i each observation belongs to.
- Lesson: Care must be taken when classifying for inferences and prediction. Especially when g unknown.

Summary

- Mixture models very attractive for model heterogeneous data.
- Wide applicability.
- In most applications, it matters, to some extent, which M_i each observation belongs to.
- Lesson: Care must be taken when classifying for inferences and prediction. Especially when g unknown.
- Using allocation helps to avoid pitfalls arising from standard conditional model selection.

Summary

- Mixture models very attractive for model heterogeneous data.
- Wide applicability.
- In most applications, it matters, to some extent, which M_i each observation belongs to.
- Lesson: Care must be taken when classifying for inferences and prediction. Especially when g unknown.
- Using allocation helps to avoid pitfalls arising from standard conditional model selection.
- THANK YOU.