An introduction to mixture modelling for unsupervised clustering

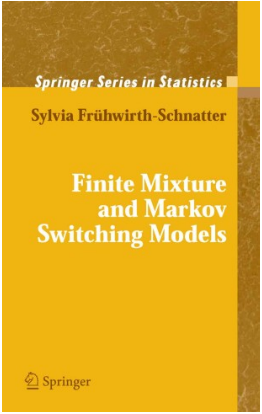
Mini-tutorial

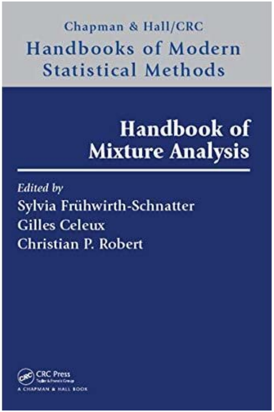
Nicole M White

Australian Centre for Health Services Innovation (AusHSI)
Queensland University of Technology

July 5, 2021

Further reading

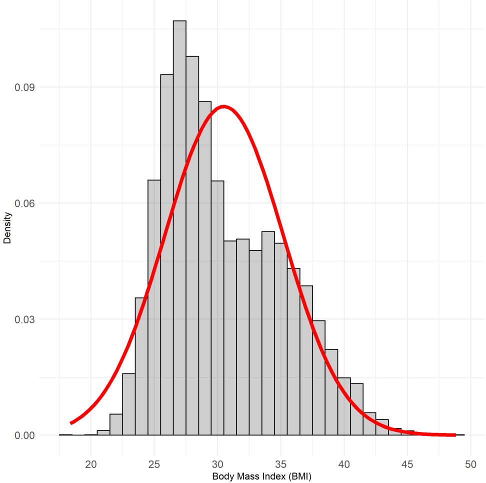




https://github.com/nicolemwhite/anzsc-mixture-modelling

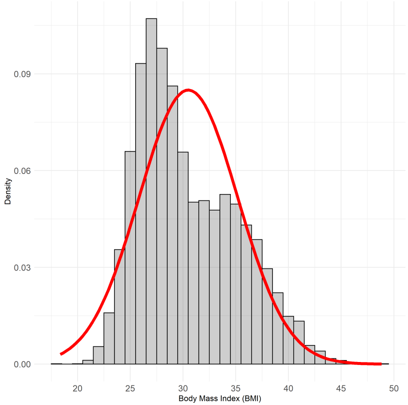
A motivating example

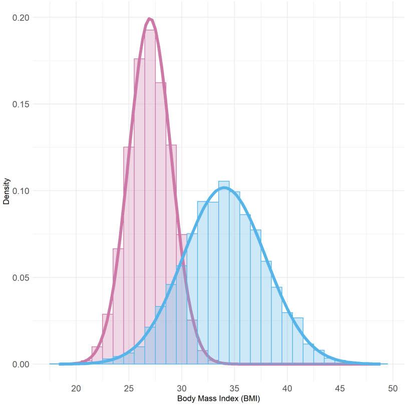
Distribution of body mass index (BMI) for 10,000 participants.



A motivating example

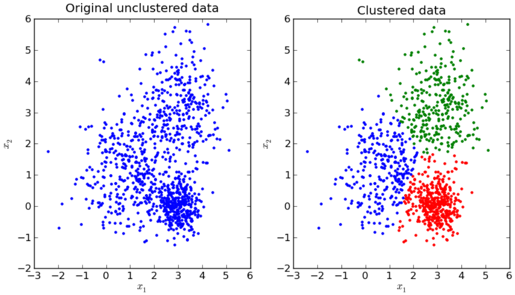
Distribution of body mass index (BMI) for 10,000 participants





Defining clustering

Unsupervised clustering $\leftrightarrow$ Identifying subgroups

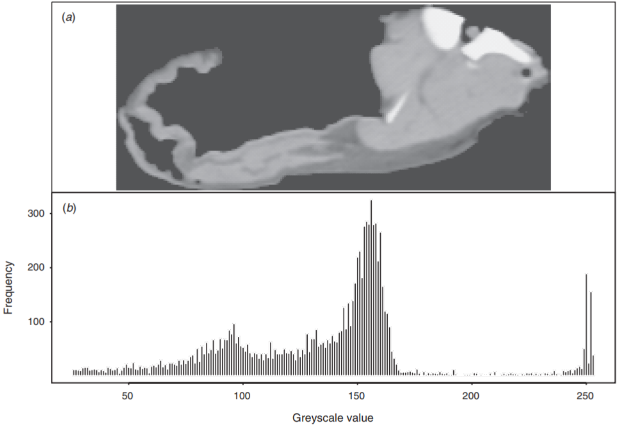


Common approaches:

- Hierarchical clustering, K-means
- Mixture models

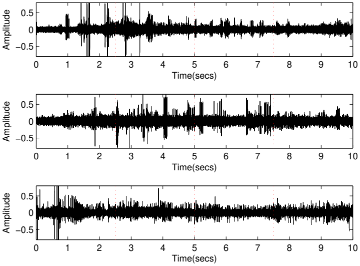
Image source: `https://tinyurl.com/59dbx8u7`

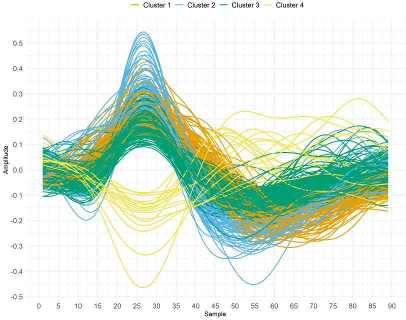
Examples of clustering using mixture models: Image classification



Alston et al (2005) DOI: 10.1071/AR04211

Examples of clustering using mixture models: Spike sorting





Mixture model ingredients

Data are drawn from a *convex combination of components* For K groups/clusters:

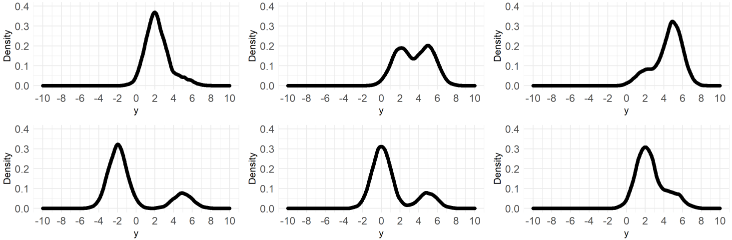
$$
\begin{aligned}
p(y) &= \eta_1 f(y|\boldsymbol{\theta}_1) + \ldots + \eta_K f(y|\boldsymbol{\theta}_K) \\
&= \sum_{k=1}^{K} \eta_k f(y|\boldsymbol{\theta}_k)
\end{aligned}
$$

Unknown parameters: $\boldsymbol{\nu} = (\boldsymbol{\eta}, \boldsymbol{\theta})$

- $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$: Mixture weights; $\sum_{k=1}^{K} \eta_k = 1$
- $f(y|\boldsymbol{\theta}_k)$: k^{th} Mixture component; same parametric family

A simple 2-component mixture model

$$y_i \sim \eta_1 \mathcal{N}(\mu_1, 1) + \eta_2 \mathcal{N}(\mu_2, 1)$$



Mixture model examples

General formulation:

$$p(y_i) = \sum_{k=1}^{K} \eta_k f(y_i|\boldsymbol{\theta}_k)$$

Latent class analysis (J items)

$$f(\mathrm{y}_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{J} f(y_{ij}|\theta_{jk})$$

Latent class regression: $\eta_k \to \eta_k(\mathrm{x}_i)$

$$\eta_k(\mathrm{x}_i) = \frac{exp\left(\mathrm{x}_i^T \beta_k\right)}{\sum_{l=1}^{K} exp\left(\mathrm{x}_i^T \beta_l\right)}$$

Mixture model examples

Focus of mini-tutorial: cross-sectional data

- Finite mixture model
- Dirichlet Process mixture model
- Profile regression

Bayesian approaches to inference: Markov chain Monte Carlo (MCMC)

1. Finite mixture models
2. Dirichlet Process Mixture models
3. Profile regression

Finite mixture model: Setup

Assume:

- K is fixed *a priori*
- Each observation has a probability of belonging to components $1, \ldots, K$

Likelihood for $\boldsymbol{y} = (y_1, \ldots, y_n)$

$$p(\boldsymbol{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \eta_k f(y_i | \boldsymbol{\theta}_k)$$

Aim is to learn $\boldsymbol{\nu} = (\boldsymbol{\theta}_{1,\ldots,K}, \boldsymbol{\eta}_{1,\ldots,K})$

Finite Mixture Model: Setup

Likelihood:

$$p(\boldsymbol{y}|\boldsymbol{\nu}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \eta_k f(y_i|\boldsymbol{\theta}_k)$$

Priors:

$$(\eta_1, \ldots, \eta_K) \sim \mathcal{D}(\gamma_1, \ldots, \gamma_K)$$
$$\theta_k \sim p(\theta_k|\delta)$$

How to estimate when membership of y_i to components $1, \ldots, K$ is not known?

Finite Mixture Model: Estimation

Enter data augmentation! (Tanner Wong, 1987; *JASA*)
The idea:

- Introduce z_i = cluster membership for y_i and treat as missing data

$$
\begin{aligned}
p(y_i|\boldsymbol{\nu}) &= \sum_{k=1}^{K} p(y_i|z_i = k, \boldsymbol{\nu}) Pr(z_i = k|\boldsymbol{\nu}) \\
Pr(z_i = k|\boldsymbol{\nu}) &= \eta_k \\
p(y_i|z_i = k, \boldsymbol{\nu}) &= f(y_i|\theta_k)
\end{aligned}
$$

- Inference on z_i provides information on clustering

Finite Mixture Model: Estimation by MCMC

1. Sample $\boldsymbol{z}$ (Bayes' rule)

$$Pr(z_i = k|y_i, \boldsymbol{\nu}) = \frac{\eta_k f(y_i|\boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \eta_j f(y_i|\boldsymbol{\theta}_j)}$$

$$z_i \sim MN(1, Pr(z_i = 1|y_i, \boldsymbol{\nu}), \ldots, Pr(z_i = K|y_i, \boldsymbol{\nu}))$$

2. Conditional on $\boldsymbol{z}$: Update $\eta_1, \ldots, \eta_K$

$$\eta_1, \ldots, \eta_K \sim \mathcal{D}(\delta_1 + n_1, \ldots, \delta_K + n_K)$$

3. Conditional on $\boldsymbol{z}$: Update $\theta_1, \ldots, \theta_K$

$$\theta_k \sim p(\boldsymbol{\theta}_k|\delta) \prod_{i:z_i=k} f(y_i|\boldsymbol{\theta}_k)$$

Finite Mixture Model: Estimation by MCMC

Available approaches in R:

- `R2OpenBUGS` (see `fmm_BUGS.R`)
- `bayesmix`

Or code from scratch:

- see `fmm_mvn.R` for Multivariate Normal example

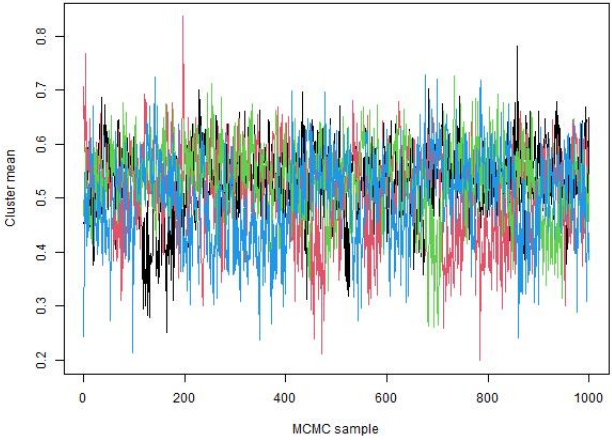
Label switching

Issue arises as likelihood is invariant to permutations of k e.g. $K = 3$

$$
\begin{aligned}
p(y_i|\boldsymbol{\nu}) &= \eta_1 p(y_i|\boldsymbol{\theta}_1) + \eta_2 p(y_i|\boldsymbol{\theta}_2) + \eta_3 p(y_i|\boldsymbol{\theta}_3) \\
&= \eta_3 p(y_i|\boldsymbol{\theta}_3) + \eta_2 p(y_i|\boldsymbol{\theta}_2) + \eta_1 p(y_i|\boldsymbol{\theta}_1) \\
&= \eta_2 p(y_i|\boldsymbol{\theta}_2) + \eta_3 p(y_i|\boldsymbol{\theta}_3) + \eta_1 p(y_i|\boldsymbol{\theta}_1)
\end{aligned}
$$

When sampling z_i, components can be relabelled $\rightarrow$ affects clustering inference

Label switching example



Label switching: Possible solutions

Prior constraints (not a good idea):

$$\eta_1 < \eta_2 < \ldots < \eta_K$$

Relabelling algorithms:

- Loss functions: minimise over all MCMC samples of z (Stephens, 2000)
- MAP estimate $\widehat{z}$ as 'pivot' (Marin et. al, 2005)

Label switching: Possible solutions

Similarity matrix, **S**:

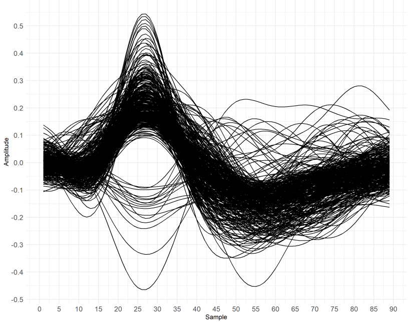
$$S_{ii'}^{(d)} = \begin{cases} 1 & \text{if } z_i^{(d)} = z_{i'}^{(d)} \\ 0 & \text{otherwise.} \end{cases}$$

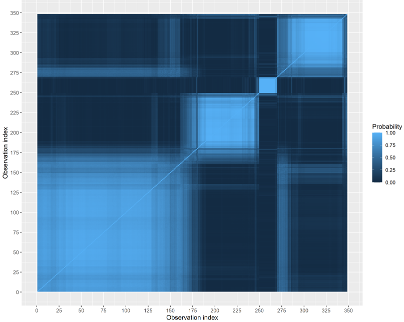
$$\overline{S} = \frac{1}{D} \sum_{d=1}^{D} S^{(d)}$$

R packages: `mcclust`, `label.switching`

Label switching

Example: Spike sorting, $K = 4$ clusters





Choosing K

Common information criteria:

- Akaike's Information Criterion (AIC)

$$AIC_K = -\log p(\mathrm{y}|\boldsymbol{\eta}^*, \boldsymbol{\theta}^*) + 2p_k$$

- Bayesian Information Criterion (BIC)

$$BIC_K = -\log p(\mathrm{y}|\boldsymbol{\eta}^*, \boldsymbol{\theta}^*) + p_k \log n$$

- Deviance Information Criterion (DIC)

$$DIC_K = -4E_{\boldsymbol{\eta},\boldsymbol{\theta}|\mathrm{y}}\left[\log p(\mathrm{y}|\boldsymbol{\eta}, \boldsymbol{\theta})\right] + 2\log f(\mathrm{y})$$

$$f(\mathrm{y}) = \prod_{i=1}^{n} \frac{1}{D} \sum_{d=1}^{D} \sum_{k=1}^{K} \eta_k^{(d)} f(y_i|\theta_k^{(d)})$$

1. Finite mixture models
2. Dirichlet Process Mixture models
3. Profile regression

Dirichlet Process mixture model: Motivation

General formulation:

$$p(y_i|\boldsymbol{\nu}) = \sum_{k=1}^{K} \eta_k f(y_i|\boldsymbol{\theta}_k)$$

In a finite mixture - assume K as fixed $\rightarrow$ model comparison problem

Alternative: Infer K as part of modelling

$$p(y_i|\boldsymbol{\nu}) = \sum_{k\geq 1} \eta_k f(y_i|\boldsymbol{\theta}_k)$$

Dirichlet Process mixture model: Setup

- Nonparametric approach to mixture modelling
- Does not estimate K directly; focus on clustering of individual parameters, θ_i

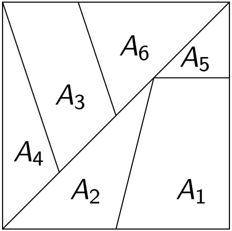
Dirichlet process (DP) prior:

$$\begin{aligned} G &\sim DP(\alpha, G_0) \\ G_0 &= E(G); \qquad \text{Base distribution} \\ \alpha &> 0; \qquad \text{Concentration parameter} \end{aligned}$$

Dirichlet Process mixture model: Setup

Each draw from a DP is itself a distribution:

$$G(A_1), \ldots, G(A_k)|\alpha, G_0 \sim D(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$$



"Discreteness property": multiple draws from $DP(\alpha, G_0)$ can take the same value; induces clustering behaviour

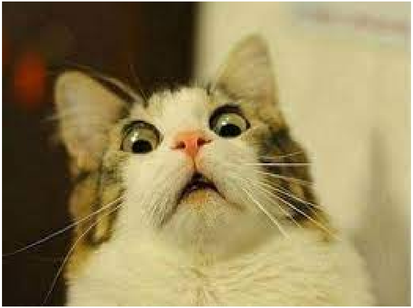
Dirichlet Process mixture model: Setup

DP as prior within mixture setting:

$$
\begin{aligned}
y_i|\boldsymbol{\theta}_i &\sim p(y_i|\boldsymbol{\theta}_i) \\
\boldsymbol{\theta}_i|G &\sim G \\
G &\sim DP(\alpha, G_0)
\end{aligned}
$$

- G is the mixing distribution
- G_0 prior distribution on unknown components
- α controls variation around G_0

Dirichlet Process mixture model: Estimation



How to sample from $DP(\alpha, G_0)$?!
What happened to η_k?!

- Stick breaking representation
- Pòlya Urn scheme/Chinese restaurant process

Stick breaking representation

G replaced by an infinite sum of weighted point masses:

$$\begin{aligned} G &\sim DP(\alpha, G_0) \\ G &= \sum_{k=1}^{\infty} \eta_k \delta_{\theta_k} \\ \theta_k &\sim G_0. \end{aligned} \tag{1}$$

$\eta_k; k = 1, \ldots$ are the "stick breaking weights".
Weights are drawn sequentially:

$$\begin{aligned} w_k|\alpha &\sim Beta(1, \alpha) \\ \eta_1 &= w_1 \\ \eta_k &= w_k \prod_{l=1}^{k-1} (1 - w_l). \end{aligned}$$

Stick breaking process

1

- Draw $w_1 \sim Beta(1, \alpha)$ and set $\eta_1 = w_1$

w_1	$1 - w_1$

- Draw $w_2 \sim Beta(1, \alpha)$ and compute $\eta_2 = w_2(1 - w_1)$

w_1	w_2	$(1 - w_2)(1 - w_1)$

- Draw $w_3 \sim Beta(1, \alpha)$ and compute $\eta_3 = w_3(1 - w_1)(1 - w_2)$

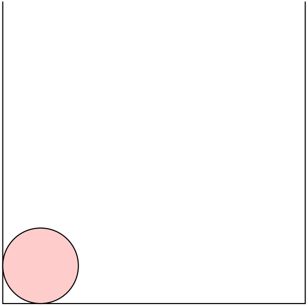
w_1	w_2	w_3	$(1 - w_3)(1 - w_2)(1 - w_1)$

As $K \to \infty$

$$p(y_i|\boldsymbol{\nu}) = \sum_{k=1}^{K} \eta_k p(y_i|\boldsymbol{\theta}_k)$$

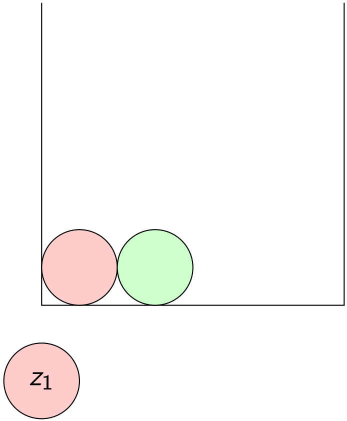
The Pòlya Urn scheme

Good analogy for implied clustering behaviour
Begin with α red balls in an urn:



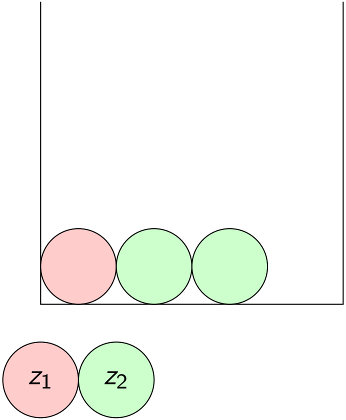
The Pòlya Urn scheme

If a red ball is drawn, record colour and replace with a ball of a new colour:

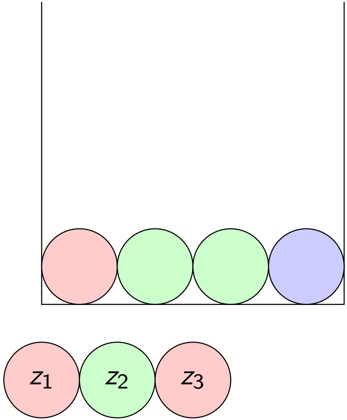


The Pòlya Urn scheme

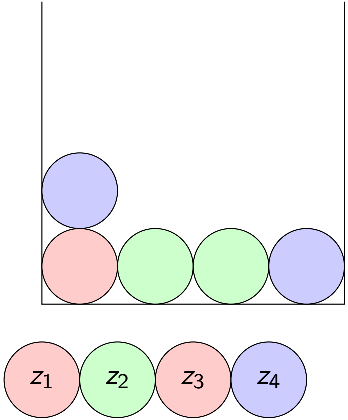
If a non-red ball is drawn, record colour and replace with a ball of the same colour:



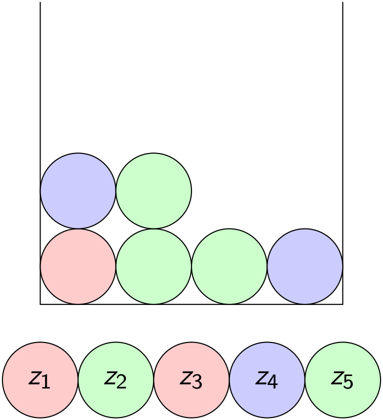
The Pòlya Urn scheme



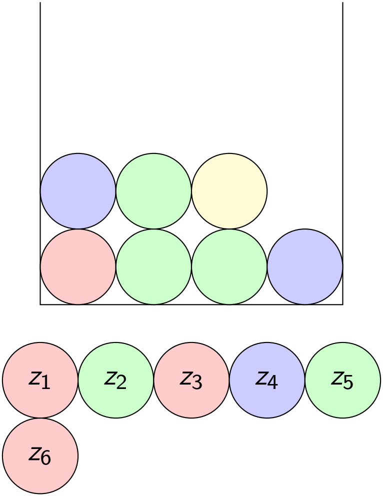
The Pòlya Urn scheme



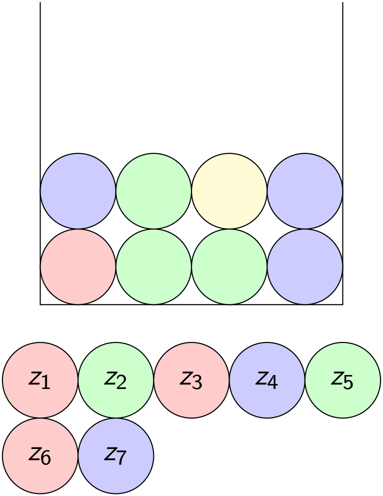
The Pòlya Urn scheme



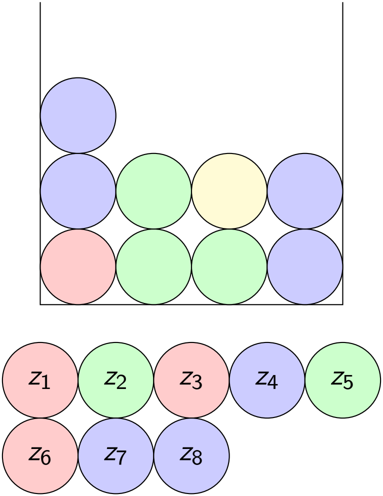
The Pòlya Urn scheme



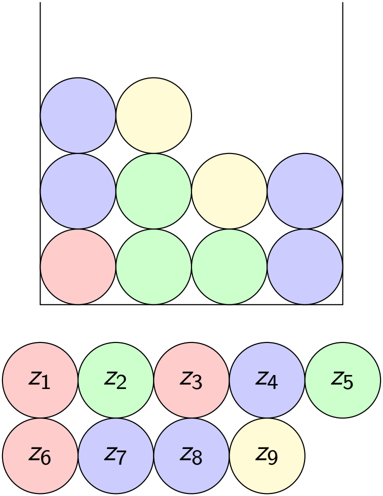
The Pòlya Urn scheme



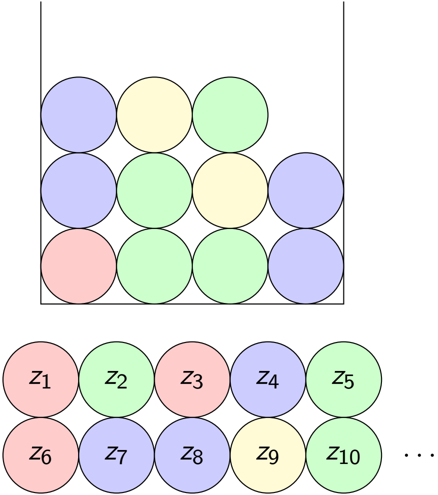
The Pòlya Urn scheme



The Pòlya Urn scheme



The Pòlya Urn scheme



- The more often a colour is drawn, the more likely it is to

Dirichlet Process mixture model: Estimation

R packages

- `dirichletprocess`
- https://rdrr.io/cran/NPflow/man/DPMGibbsN$_SeqPrior.html$`PReMiuM`

1. Finite mixture models
2. Dirichlet Process Mixture models
3. Profile regression