Clustering Categorical Data

Louise McMillan

Acknowledgements

Shirley Pledger, Daniel Fernández, Richard Arnold, Ivy Liu, Eleni Matechou, and thanks to Murray Efford for help with clustglm

Rachel Fewster and Emma Carroll for population genetics collaborations

Adam Glucksman for work on genetic clustering

Clustering

Clustering ordinal data

Common example is survey with questions answered on a scale from 1 to q

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

Row clustering

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

Find groups of individuals with similar responses

Column clustering

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

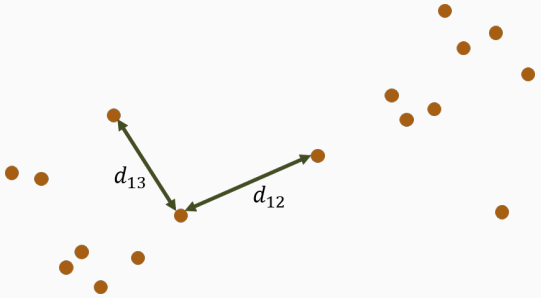
Find groups of variables that have similar responses

Biclustering

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

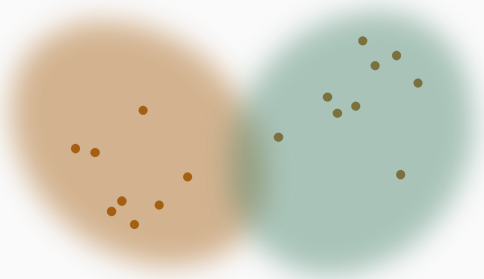
Find groups of individuals and variables simultaneously

Distance-based clustering



Assigns individuals to clusters, but requires distance metric

Model-based clustering



Fits statistical model to each cluster, providing inference about cluster feature

TL:DR

The key difficulties of clustering categorical data:

- It contains less information than numerical data, though ordinal data has slightly more information than categorical data
- There are far fewer methods designed to handle it

Approach 1: Treat the ordinal data as numerical

How do we handle ordinal data?

Name	Age	Severity	Exercise Level	Diastolic blood pressure
Wen	23	1 – Mild	Cycling	82
Mirai	57	1 – Mild	Running	98
An	51	3 – Severe	Cycling	112
Max	43	2 – Moderate	Rowing	72

A common approach is to convert ordinal data to numerical labels. . .

How do we handle ordinal data?

Name	Age	Severity	Exercise Level	Diastolic blood pressure
Wen	23	1	Cycling	82
Mirai	57	1	Running	98
An	51	3	Cycling	112
Max	43	2	Rowing	72

. . . and then treat the numerical labels as numerical data

Essentially, we pretend it contains more information than it actually does

Example: simulated ordinal data

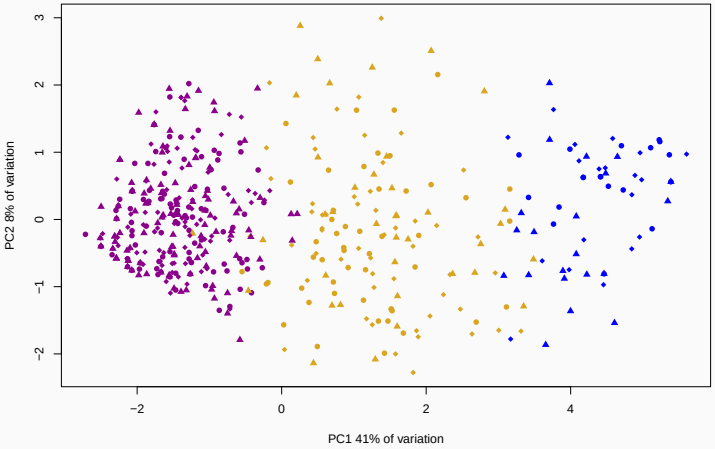
This is a constructed example of ordinal data, with 500 observations and 10 variables. Each variable has 3, 4 or 5 categories.

The observations are in 3 clusters, in proportions 10%, 30% and 60%

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10
ID1	5	5	4	3	3	3	5	3	2	1
ID2	5	3	3	3	3	3	5	3	3	3
ID3	4	3	4	2	3	3	4	2	3	3
ID4	5	4	3	2	2	1	4	1	2	3
ID5	1	2	1	4	1	1	1	1	1	1
ID6	2	5	5	2	1	3	5	2	1	3

Example 1: simulated ordinal data

We can plot the first two principal components after standardizing the data:



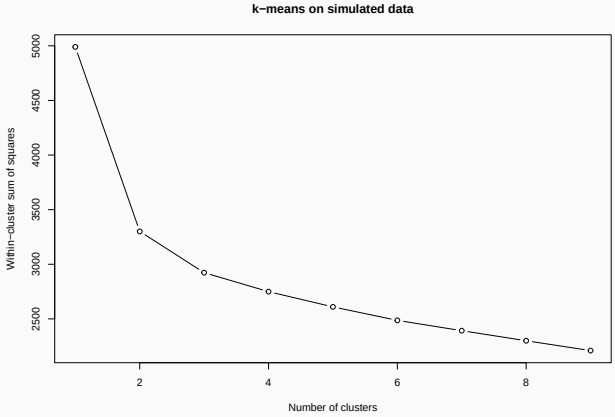
k-means

k-means is the most common clustering method, and it is based on distance and variance

It iteratively reallocates observations to clusters to minimise the within-cluster variances

k-means

I tried 2 to 9 clusters, using 50 starts and up to 1000 iterations, and compared the within-cluster sums of squares:



k-means

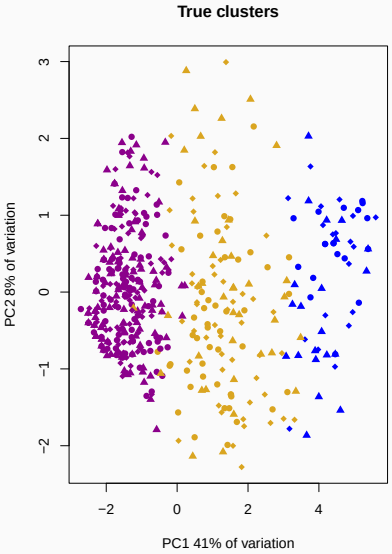
We can use the Adjusted Rand Index to compare the cluster allocations to the true ones without being affected by label switching

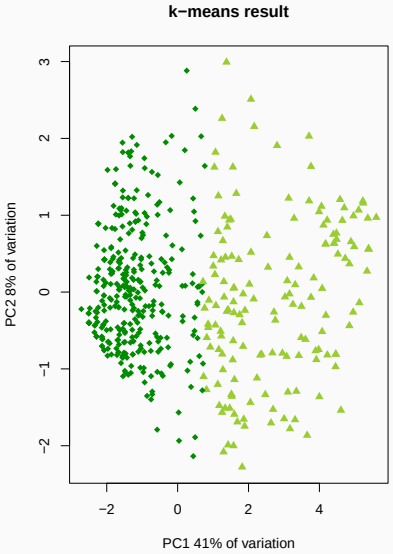
The Adjusted Rand Index theoretically ranges from -1 to 1, where 1 indicates a perfectly matched pair of cluster allocations:

	kmeans
ARI	0.6278382

This result is pretty good, and it's certainly better than chance, which corresponds to 0

k-means


True clusters
PC2 8% of variation
PC1 41% of variation
3
2
1
0
−1
−2
−2
0
2
4


k−means result
PC2 8% of variation
PC1 41% of variation
3
2
1
0
−1
−2
−2
0
2
4

Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are a common model-based clustering method for numerical data

The data are assumed to arise from a mixture of multivariate Gaussian distributions, and you specify the covariance structures when fitting the model

I will test the most flexible covariance structures, in which the variances and correlations are allowed to vary amongst dimensions and amongst clusters

Gaussian Mixture Models

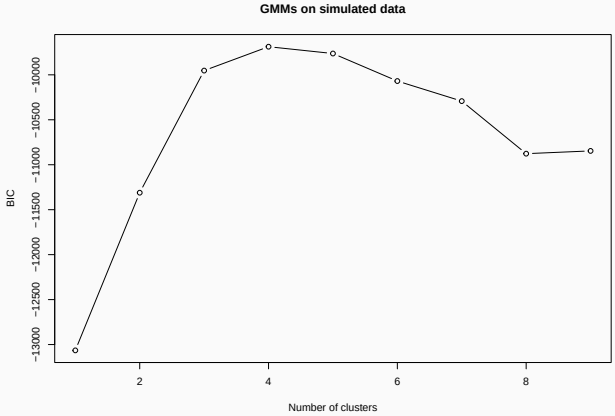
The model is fitted using the EM algorithm, and the `mclust` GMM functionality also uses hierarchical clustering to find the starting points for the EM algorithm

Scrucca L., Fraley C., Murphy T. B. and Raftery A. E. (2023) Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman & Hall/CRC, ISBN: 978-1032234953, https://mclust-org.github.io/book/

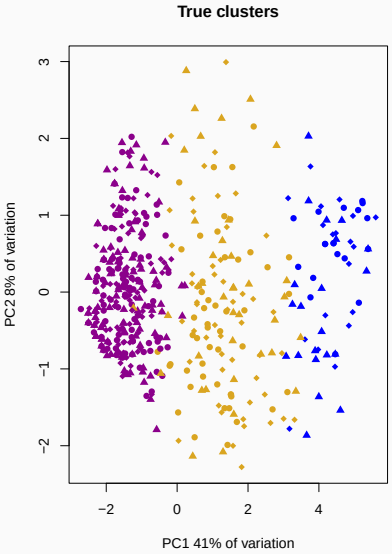
Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models, The R Journal, 8/1, pp. 289-317.

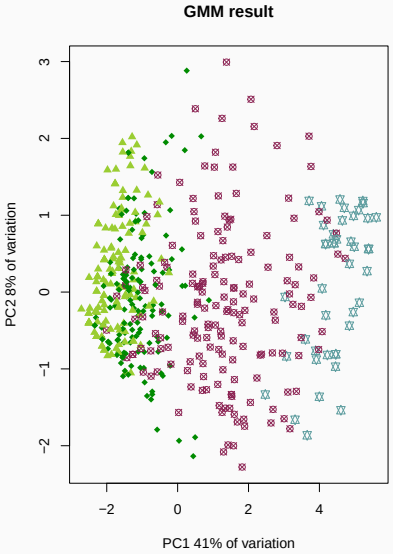
Gaussian Mixture Models

The model selection criterion is BIC, and the `mclust` form of BIC is higher for better-fitted models



Gaussian Mixture Models





Gaussian Mixture Models

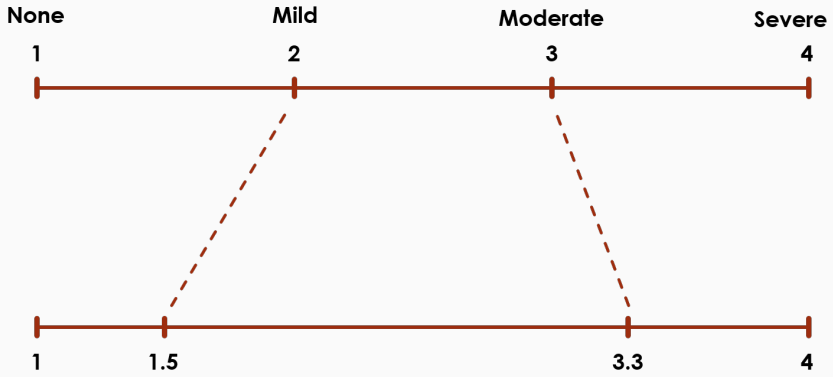
Use the Adjusted Rand Index again to check the GMM result against the true clusters, repeating the ARI for k-means for comparison:

	kmeans	GMM
ARI	0.6278382	0.3592375

This result is better than chance, but worse than the result for k-means

Approach 2: Treat the ordinal data as ordinal

The problem with numerical labels



The problem with numerical labels

The numerical labels may not accurately describe the spacing!

It would be just as valid to number the levels -54, -9, 39 and 240 as to number them 1 to 4

The only detail the ordinal nature of the data specifies is the order, not the spacing

So if we want to fit the data without making the assumption of equal spacing, we should use methods that treat the data as ordinal

CUB models

One type of model, designed originally by Domenico Piccolo, is specifically designed for ordinal data that arise from questionnaires

Combination of Uniform and Binomial models assume that each individual response to a particular question, and from a particular observation, has a component which is the true opinion, modelled as a binomial, and another component which is the uncertainty, modelled as a uniform

The original models were developed for regressing a single response variable on covariates

CUB models

Matteo Ventura has been working with Julian Jacques, Paola Zuccolotto and Domenico Piccolo on clustering CUB models

The overall model is a nested mixture of mixtures

It has already proved useful for interpreting patterns of opinions from survey data

However, the code is not yet available as an R package

Ventura, M., Jacques, J. & Zuccolotto, P. "Model-Based Clustering of Multivariate Rating Data Accounting for Feeling and Uncertainty." *Journal of Classification* (2025).

clustMD

The clustMD package was developed to cluster mixed-type data

It can handle a mixture of continuous, ordinal and nominal data

It assumes that there is a latent multivariate Gaussian underpinning all of the data types

It requires more rows than columns for row clustering

McParland, D., Gormley, I.C. Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification* (2016) 10, 155–169.
https://doi.org/10.1007/s11634-016-0238-x

Binary Ordinal Search

Binary Ordinal Search is a method proposed by Biernacki and Jacques that uses ordinal-specific models

This method treats ordinal data as originating from a binary search process

The clustering algorithm is implemented in the `ordinalClust` package

It can cluster observations (rows) or variables (columns), or both at once

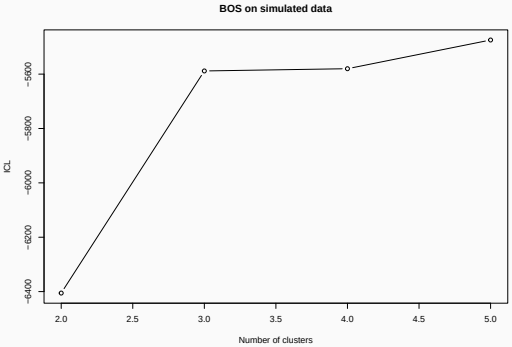
Biernacki, C., & Jacques, J. "Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm." *Statistics and Computing* (2016) 26(5), 929-943.

Binary Ordinal Search

For such a small dataset, it is difficult to fit more than 5 clusters, and we first have to rearrange the variables so that all of the ones with the same number of levels are gathered together in contiguous columns

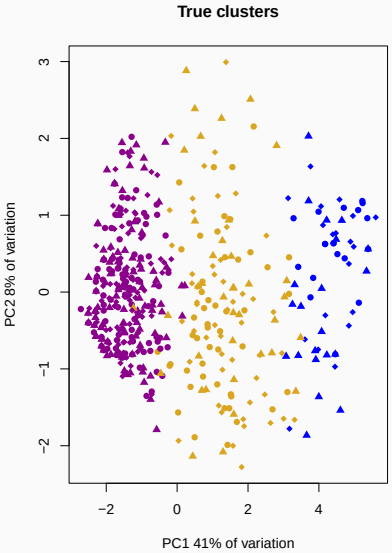
The algorithm resets the C++ seed every time it runs, so we'll retry the algorithm multiple times, in order to avoid spurious solutions:

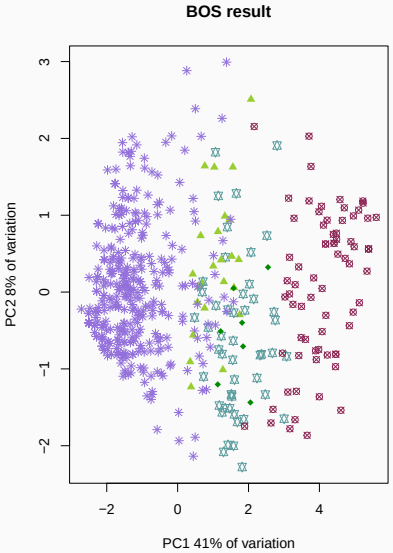
Binary Ordinal Search



This is a model-based approach, like GMMs, so model selection is based on the likelihood and on the integrated complete likelihood (ICL), and the higher values indicate a better fit, so 5 is the best number of clusters

Binary Ordinal Search





Binary Ordinal Search

Use the Adjusted Rand Index again to check the GMM result against the true clusters, repeating the ARI for k-means for comparison:

	kmeans	GMM	BOS
ARI	0.6278382	0.3592375	0.6369881

clustord

`clustord`: github.com/vuw-clustering/clustord

https://cran.r-project.org/web/packages/clustord/index.html

Why this package?

`clustord` can handle **imbalanced data**: it can identify small clusters (e.g. 4 observations out of 200), provided they're distinct from the other clusters

`clustord` can cluster observations with **missing data**: for observations with partial responses, `clustord` calculates likelihoods based on the remaining responses, instead of dropping data or imputing values

Why this package?

`clustord` fits a specified number of clusters each time

BUT some clusters may end up with no observations assigned to them, which may indicate that a smaller number of clusters is suitable

Why this package?

Clustering models akin to regression

Data matrix of response variables (Y_{ij})

Clusters affect responses via a linear predictor η_{ij} and we can also add other effects to this model

Ordinal models

Proportional odds ordinal model:

$$\log\left(\frac{P(Y_{ij} \geq k)}{P(Y_{ij} < k)}\right) = \mu_k - \eta_{ij}$$

Ordered stereotype ordinal model:

$$\log\left(\frac{P(Y_{ij} = k)}{P(Y_{ij} = 1)}\right) = \mu_k + \phi_k \eta_{ij}$$

Note that both models use the same linear predictor term η_{ij}

Row clustering

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

$$\eta_{ij} = \alpha_r \text{ for } i \in r$$

This structure assumes that each observation's responses are purely dependent on which observation cluster it is in

Column clustering

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

$$\eta_{ij} = \beta_c \text{ for } j \in c$$

This structure assumes that each variable's responses are purely dependent on which variable cluster it is in

Biclustering

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

$$\eta_{ij} = \alpha_r + \beta_c \text{ for } i \in r \text{ and } j \in c$$

This structure assumes that observation clusters and variable clusters both have effects that change the response

Row clustering with column effects

Name	Q1	Q2	Q3	Q4	Q5	Q6
Wen	3	3	2	3	3	3
Mirai	1	2	3	1	3	2
An	2	2	2	1	3	2
Max	2	1	1	1	2	1

$$\eta_{ij} = \alpha_r + \beta_j \text{ for } i \in r$$

This structure assumes that each observation's responses are partially dependent on which observation cluster it is in, but that the responses also vary by variable

Accounts for some variables having unusual response patterns

Row clustering with row covariates

Name	Q1	Q2	Q3	Q4	Q5	Q6	Age
Wen	3	3	2	3	3	3	29
Mirai	1	2	3	1	3	2	51
An	2	2	2	1	3	2	59
Max	2	1	1	1	2	1	33

$$\eta_{ij} = \alpha_r + \boldsymbol{\delta}^T \boldsymbol{x}_i \text{ for } i \in r$$

This structure incorporates covariates, additional information about observations, that affects their responses

Ordinal models

`clustord` is designed to cluster ordinal data with a **fixed number** of levels, but it works on data with a varying number of levels, by treating all the columns as if they have the maximum number of levels out of all of them

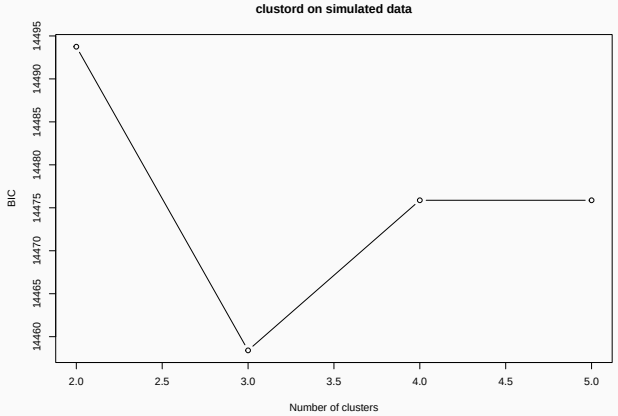
This works better if you incorporate column effects, because those can indirectly model the variables with fewer response levels

clustord

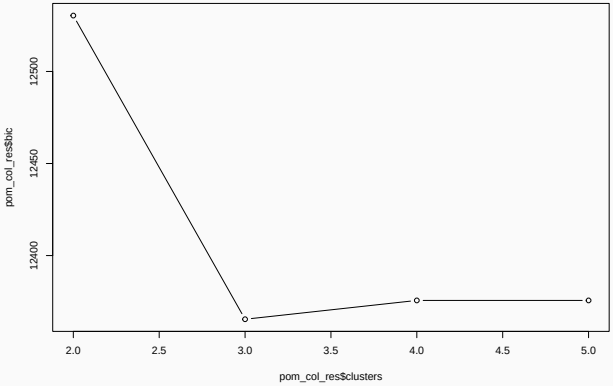
First try fitting the simplest `clustord` model, with only the row clusters and no individual column effects: `Y ~ ROWCLUST`, and try 2 to 5 clusters

I will use the proportional odds model (POM), which is more restrictive than the ordered stereotype model (OSM)

The model selection criterion is BIC, but this form of BIC is lower for better-fitted models

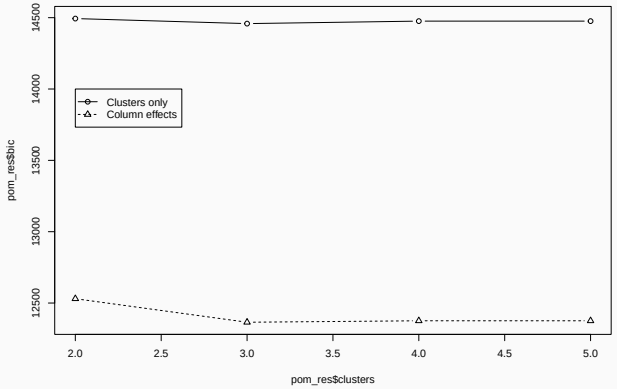


Then try the model with individual column effects



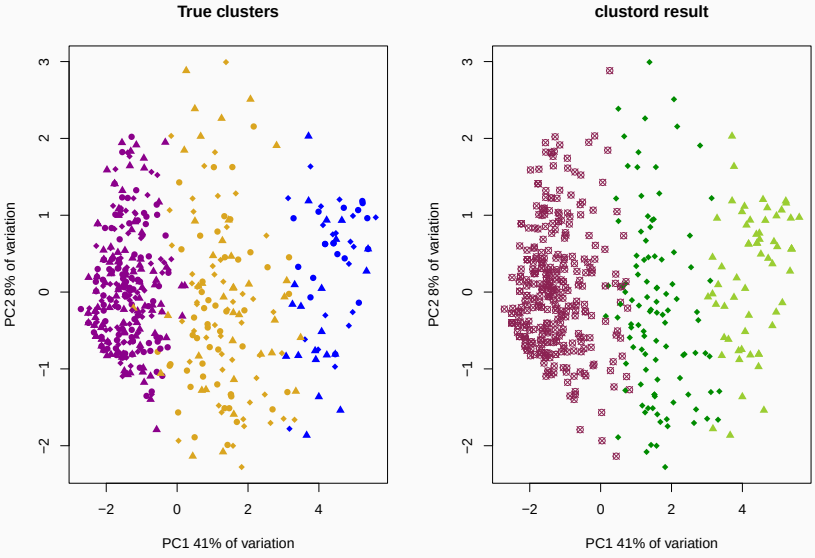
Again, the best number of clusters is 3

Compare the two types of models



The model with column effects is clearly a better fit

clustord


True clusters
clustord result
PC2 8% of variation
PC1 41% of variation
3
2
1
0
−1
−2
−2
0
2
4

Use Adjusted Rand Index again to assess the performance

	kmeans	GMM	BOS	clustord
ARI	0.6278382	0.3592375	0.6369881	0.7399687

Simulating ordinal data

I am working on a full simulation study comparison of clustering methods, including `clustord`, and a real data study using the Quality of Life dataset analysed in the past by `ordinalClust`

This is just one result, a single demonstration, but I did make life a bit harder by using OSM, which is more flexible than POM, to simulate the ordinal data, and by having varying numbers of levels in the variables

Simulating ordinal data

I simulated a matrix of responses, Y_{ij}, with q_j levels per variable, indexed $k = 1, \ldots, q_j$ and $q_j \in \{3, 4, 5\}$

I used the ordered stereotype model with cluster effects $\boldsymbol{\alpha} = (-3, 0, 3)$ and additional effects of individual columns $\beta_j \in [-1, 1]$

For observation/row i in cluster g

$$\log\left(\frac{P(Y_{ij} = k)}{P(Y_{ij} = 1)}\right) = \mu_{jk} + \phi_{jk}(\alpha_g + \beta_j)$$

Simulating ordinal data

It is also possible to simulate ordinal data by simulating data from normal distributions and selecting cut points to split the values up into the different ordinal levels

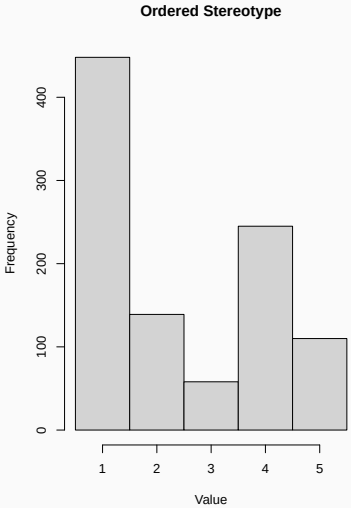
However, in order to obtain variables with the full flexibility of more flexible ordinal models, you have to play around with the cut points, not just vary the shape of the latent normal distribution

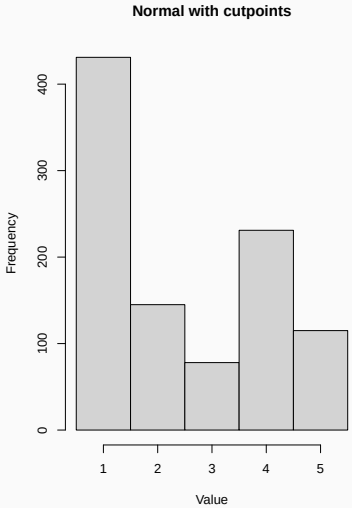
Methods that treat ordinal data as normal with cut points often do not fit the cut points in a data-driven way

Simulating ordinal data

The following two distributions are 1000 data points simulated from an ordered stereotype model with random μ_k and ϕ_k values, $\alpha = -2$ and $\beta = 1$, and 1000 data points simulated from a standard normal distribution with cut points -0.2, 0.2, 0.4 and 1.2

Simulating ordinal data





Genetic clustering

DNA dataset

ID	L1.a1	L1.a2	L2.a1	L2.a2	L3.a1	L3.a2
Kai-3	128	128	112	140	136	138
Kai-4	96	128	140	140	130	136
Kai-6	92	96	140	142	120	120
Nel-1	96	96	112	140	120	138
Nel-3	128	130	112	112	124	138

Microsatellite data looks numerical but is ordinal/nominal categorical

SNP data is binary

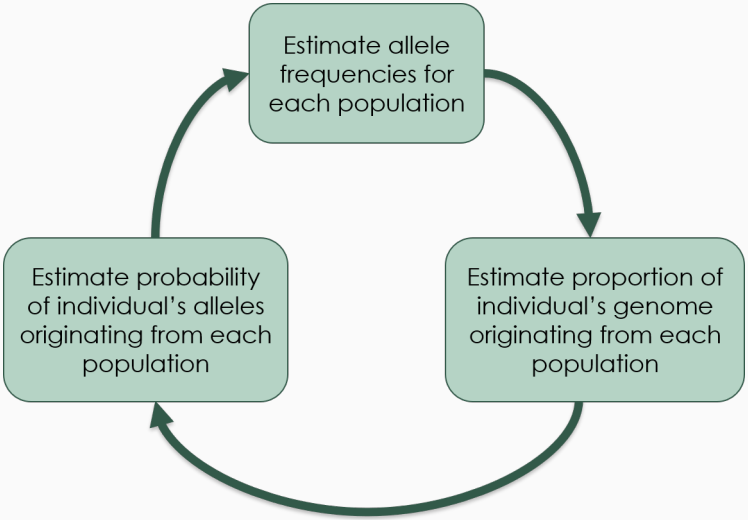
Diploid genetic data has more than one response per variable (2 copies of each chromosome)

Existing methods

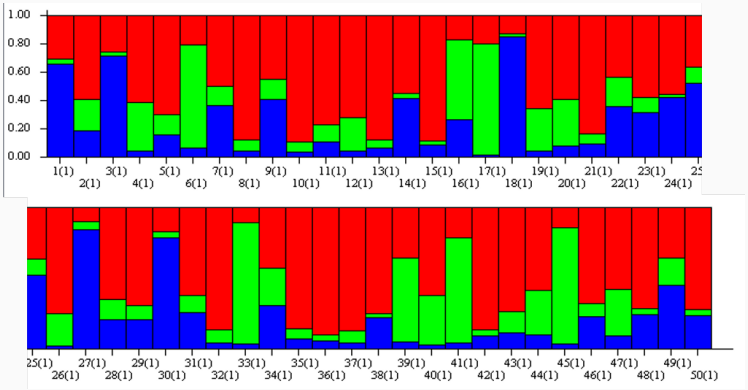
Many existing methods for performing unsupervised genetic clustering

STRUCTURE, fastSTRUCTURE, fineSTRUCTURE, ADMIXTURE, BayesAss, and more...

Bayesian method STRUCTURE



Bayesian method STRUCTURE



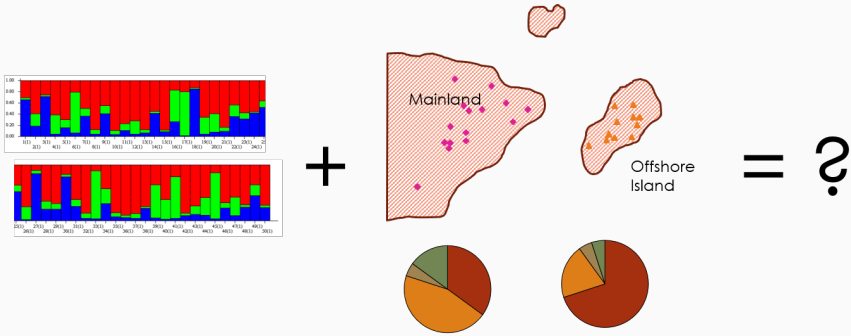
Different DNA types

ID	L1.a1	L1.a2	L2.a1	L2.a2	L3.a1	L3.a2	mtDNA
Kai-3	128	128	112	140	136	138	P
Kai-4	96	128	140	140	130	136	L
Kai-6	92	96	140	142	120	120	A
Nel-1	96	96	112	140	120	138	P
Nel-3	128	130	112	112	124	138	O

Biologists also collect mitochondrial DNA haplotypes

Ecologists use separate analyses for mitochondrial DNA

Different DNA types



Analyse nuclear DNA and mitochondrial DNA separately

Genetic and non-genetic data

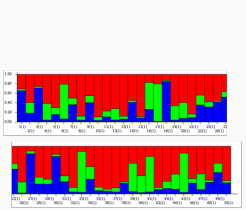
ID	L1.a1	L1.a2	L2.a1	L2.a2	L3.a1	L3.a2	mtDNA	15N
Kai-3	128	128	112	140	136	138	P	8.3
Kai-4	96	128	140	140	130	136	L	7.1
Kai-6	92	96	140	142	120	120	A	6.2
Nel-1	96	96	112	140	120	138	P	9.2
Nel-3	128	130	112	112	124	138	O	9.1

They have additional dietary isotope data, or could collect other environmental or behavioural covariates. . .

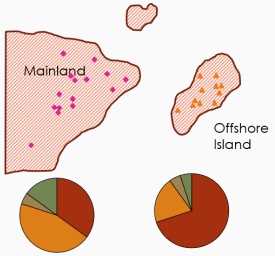
Genetic and non-genetic data

https://tohoravoyages.ac.nz/track-the-2021-tohora/

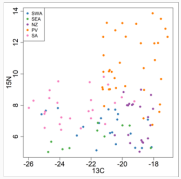
Genetic and non-genetic data



+



+



=?

Current approach: analyse nuclear DNA and mitochondrial DNA separately

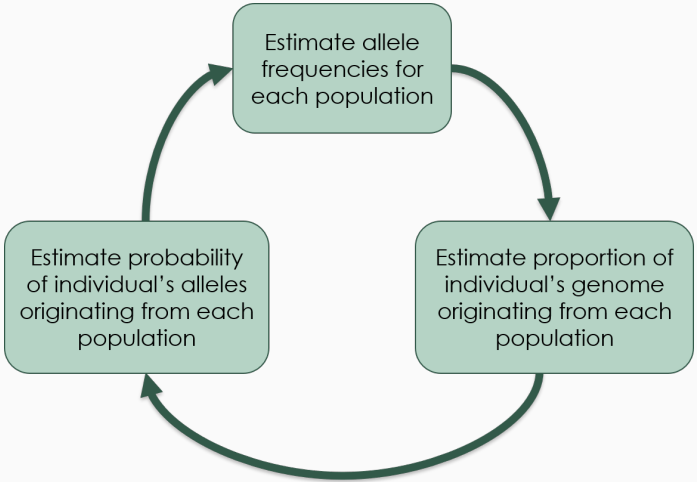
Modified Bayesian clustering

Adam Glucksman is working on combining nuclear and mitochondrial DNA in the same clustering analysis

Mitochondrial DNA provides extra evidence about which population the animals may have originated in

Modified Bayesian clustering

Can we modify this model to incorporate isotope data?



Modified Bayesian clustering

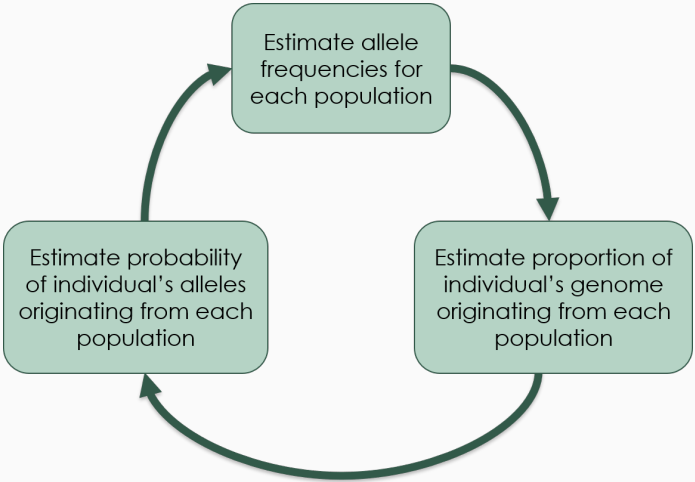
We also want to incorporate isotope data, which captures where the individual feeds

Feeding ground preferences are driven by maternal inheritance, and mitochondrial DNA is also maternally inherited

So the individual's feeding ground preferences provide extra information about their mother's population of origin

Modified Bayesian clustering

Can we modify this model to incorporate isotope data?



Thanks!