Riffle: an R Package for Nonmetric Clustering

Geoffrey B. Matthews and Robin A. Matthews

Western Washington University

Bellingham, WA, USA

Problems for Multivariate Data Analysis

- Censored data.
  - Tied ranks and reduced variance when “<5” ⇒ “5”.
  - Systematic bias when omitted.
- Missing data.
  - Omit entire row when one variable column is missing?
- Noisy, “useless” parameters.
  - Measured anyway.
  - Can be unrelated to major patterns.

- Dissimilar data types
  - Chemical
    - ph, alkalinity
  - Physical
    - temperature, percent canopy cover, sediment size, land use classes
  - Biological
    - chlorophyll, sex (male, female, juvenile)
    - rare species (counts 1-2)
    - common species (counts 10,000-100,000)

Riffle

Matthews & Hearne, *IEEE PAMI,* 1991

A clustering algorithm:
- group similar points into clusters.

A nonmetric algorithm:
- uses only order statistics for continuous data
- can handle both continuous and categorical data together

Uses variables independently:
- ignores scattered missing values
- uses incommensurable variables without normalizing

Proportional Reduction in Error

- Measuring Predictability for Categorical Variables

	red	green	blue	Errors
A	5	8	2	7
B	2	3	9	5
C	8	1	0	1
Totals	15	12	11	13

Errors predicting (red, green, blue) *a priori*: $12+11=23$

Errors predicting (red, green, blue) *given* (A, B, C): $7+5+1=13$

Proportional reduction in error: $\frac{23-13}{23}=\frac{10}{23}$

- More meaningful and robust than, *e.g.*, χ^2

Proportional Reduction in Error

Independent variables:

	red	green	blue	Errors
A	6	3	9	9
B	4	2	6	6
C	2	1	3	3
Totals	12	6	18	18

$\frac{0}{18}$

Minimum:
0% reduction

Perfectly predictable variables:

	red	green	blue	Errors
A	12	0	0	0
B	0	0	18	0
C	0	6	0	0
Totals	12	6	18	0

$\frac{18}{18}$

Maximum:
100% reduction

Clustering with categorical variables

- Assign clusters to maximize predictability over other variables.

Point	Cluster
1	A
2	A
3	B
4	C
5	A
6	C
7	B
8	C
9	A
10	C
11	B
⋮	⋮

	Variable 1		
A	1	8	2
B	3	4	1
C	1	2	3

$\frac{1}{11}$

	Variable 2		
A	9	3	2
B	2	8	2
C	1	9	3

$\frac{5}{11}$

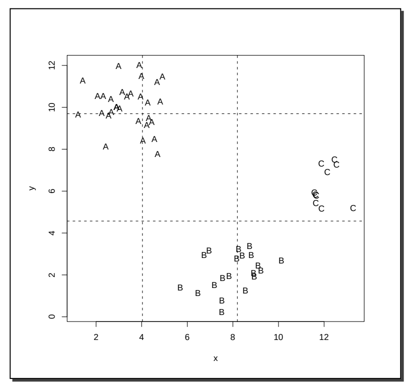
	Variable 3		
A	8	3	7
B	2	4	3
C	6	6	2

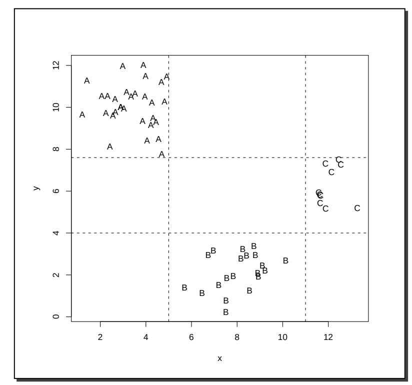
$\frac{2}{25}$

...

Handling ordered variables

- Cuts adjusted to maximize predictability of clusters





	x		
A	20	10	0
B	0	10	10
C	0	0	10

20/40

	y		
A	0	10	20
B	20	0	0
C	0	10	0

30/40

	x		
A	30	0	0
B	0	20	0
C	0	0	10

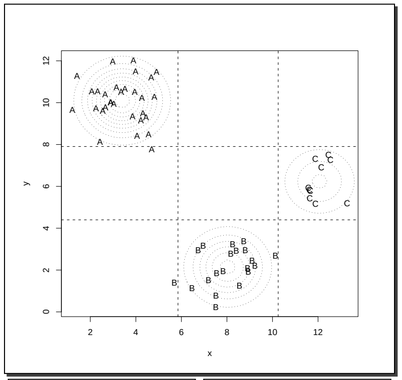
30/30

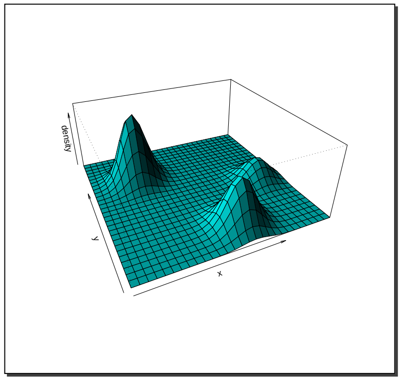
	y		
A	0	0	30
B	20	0	0
C	0	10	0

30/30

Cutting Gaussian variables

- Generate independent Gaussians from cluster statistics μ_i, σ_i
- Cut where max likelihood changes from one to another.





	x		
A	30	0	0
B	1	19	0
C	0	0	10

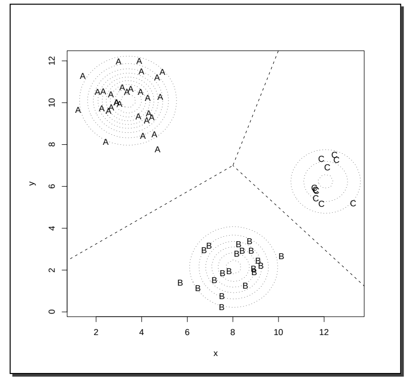
28/29

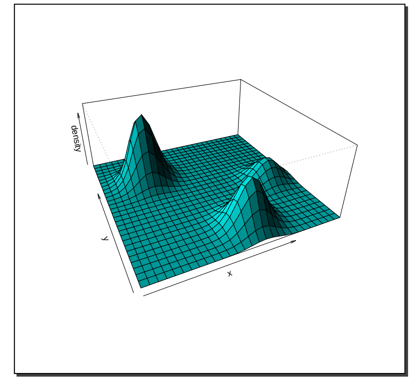
	y		
A	0	1	29
B	20	0	0
C	0	10	0

30/31

Alternative handling of Gaussian variables (EM)

- Assign to most likely group, instead of max predictability.
- Not used in Riffle.





Essential Algorithm

```
variables <- quantile.cuts(data)
clusters <- seed.clusters(variables)
score <- reduction.in.error(variables, clusters)

while (improving(score)) {
  variables <- best.cuts(variables, clusters)
  clusters <- best.clusters(variables, clusters)
  score <- reduction.in.error(variables, clusters)
}

return (clusters, variables)
```


Getting things started

To find initial cuts for variables:

- Use quantiles for cut points.
- Use quantiles for μ_i, overall σ for σ_i.

To find initial clusters, given cut variables:

- Select one point randomly as seed.
- Find other seeds by selecting points as different as possible.
- Assign each seed to a different cluster.
- Assign all other points to cluster of most similar seed.

Embellishments

- Each variable is dealt with independently.
- Each variable has a score (predictability *vs.* cluster).
- Use score to eliminate variables, or rank them in importance.
- We use this to handle the curse of dimensionality and find a small set of critical variables.
- Data reduction

Data Exploration *vs.* Confirmation

- Clustering in general is exploratory.
- Clustering data with known groups:
  - correlation between clusters and groups measures significance.
  - identifies important variables as the ones with high predictibility.
  - determine not only significance of effect, but also which variables are affected the most.
  - we have used this to chart seasonal effects.

Conclusion

- We have used **Riffle** successfully for over 10 years for ecological and toxicological data analysis.
- **Riffle** can cluster using incommensurate variables.
- **Riffle** handles censored data and missing data with few assumptions.
- **Riffle** can reduce complexity in highly multivariate datasets.
- **R** package available 2006.