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)

Rifflle
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

<table>
<thead>
<tr>
<th></th>
<th>red</th>
<th>green</th>
<th>blue</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>15</td>
<td>12</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

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., \(\chi^2\)

Rifle: an R Package for Nonmetric Clustering

Clustering with categorical variables
• Assign clusters to maximize predictability over other variables.

Handling ordered variables
• Cuts adjusted to maximize predictability of clusters

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

Rifle: an R Package for Nonmetric Clustering
Cutting Gaussian variables
- Generate independent Gaussians from cluster statistics $\mu_i, \sigma_i$
- Cut where max likelihood changes from one to another.

Essential Algorithm

```r
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 $\mu_i$, overall $\sigma$ for $\sigma_i$.

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.