Summer School in Statistics for Astronomers VI

June 9, 2010

Multivariate Analysis

James L Rosenberger

Acknowledgements:
Donald Richards
Department of Statistics
Center for Astrostatistics
Penn State University
Multivariate analysis: The statistical analysis of data containing observations on two or more variables each measured on a set of objects or cases.


63,501 objects: galaxies

http://astrostatistics.psu.edu/datasets/COMBO17.dat
<table>
<thead>
<tr>
<th>Rmag</th>
<th>mumax</th>
<th>Mcz</th>
<th>MCzml</th>
<th>chi2red</th>
<th>UjMAG</th>
<th>BjMAG</th>
<th>VjMAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.995</td>
<td>24.214</td>
<td>0.832</td>
<td>1.400</td>
<td>0.64</td>
<td>-17.67</td>
<td>-17.54</td>
<td>-17.76</td>
</tr>
<tr>
<td>25.013</td>
<td>25.303</td>
<td>0.927</td>
<td>0.864</td>
<td>0.41</td>
<td>-18.28</td>
<td>-17.86</td>
<td>-18.20</td>
</tr>
<tr>
<td>24.246</td>
<td>23.511</td>
<td>1.202</td>
<td>1.217</td>
<td>0.92</td>
<td>-19.75</td>
<td>-19.91</td>
<td>-20.41</td>
</tr>
<tr>
<td>25.203</td>
<td>24.948</td>
<td>0.912</td>
<td>0.776</td>
<td>0.39</td>
<td>-17.83</td>
<td>-17.39</td>
<td>-17.67</td>
</tr>
<tr>
<td>25.504</td>
<td>24.934</td>
<td>0.848</td>
<td>1.330</td>
<td>1.45</td>
<td>-17.69</td>
<td>-18.40</td>
<td>-19.37</td>
</tr>
<tr>
<td>23.740</td>
<td>24.609</td>
<td>0.882</td>
<td>0.877</td>
<td>0.52</td>
<td>-19.22</td>
<td>-18.11</td>
<td>-18.70</td>
</tr>
<tr>
<td>25.706</td>
<td>25.271</td>
<td>0.896</td>
<td>0.870</td>
<td>1.31</td>
<td>-17.09</td>
<td>-16.06</td>
<td>-16.23</td>
</tr>
<tr>
<td>25.139</td>
<td>25.376</td>
<td>0.930</td>
<td>0.877</td>
<td>1.84</td>
<td>-16.87</td>
<td>-16.49</td>
<td>-17.01</td>
</tr>
<tr>
<td>24.699</td>
<td>24.611</td>
<td>0.774</td>
<td>0.821</td>
<td>1.03</td>
<td>-17.67</td>
<td>-17.68</td>
<td>-17.87</td>
</tr>
<tr>
<td>24.849</td>
<td>24.264</td>
<td>0.062</td>
<td>0.055</td>
<td>0.55</td>
<td>-11.63</td>
<td>-11.15</td>
<td>-11.32</td>
</tr>
<tr>
<td>25.309</td>
<td>25.598</td>
<td>0.874</td>
<td>0.878</td>
<td>1.14</td>
<td>-17.61</td>
<td>-16.90</td>
<td>-17.58</td>
</tr>
<tr>
<td>25.219</td>
<td>25.050</td>
<td>1.109</td>
<td>1.400</td>
<td>1.76</td>
<td>-18.57</td>
<td>-18.49</td>
<td>-18.76</td>
</tr>
<tr>
<td>26.269</td>
<td>25.039</td>
<td>0.143</td>
<td>0.130</td>
<td>1.52</td>
<td>-10.95</td>
<td>-10.30</td>
<td>-11.82</td>
</tr>
<tr>
<td>23.596</td>
<td>23.885</td>
<td>0.626</td>
<td>0.680</td>
<td>0.78</td>
<td>-17.75</td>
<td>-18.21</td>
<td>-19.11</td>
</tr>
<tr>
<td>23.204</td>
<td>23.517</td>
<td>1.185</td>
<td>1.217</td>
<td>1.79</td>
<td>-20.50</td>
<td>-20.14</td>
<td>-20.30</td>
</tr>
<tr>
<td>25.161</td>
<td>25.189</td>
<td>0.921</td>
<td>0.947</td>
<td>1.68</td>
<td>-17.87</td>
<td>-16.13</td>
<td>-16.30</td>
</tr>
<tr>
<td>22.884</td>
<td>23.227</td>
<td>0.832</td>
<td>0.837</td>
<td>0.20</td>
<td>-19.81</td>
<td>-19.42</td>
<td>-19.64</td>
</tr>
<tr>
<td>24.346</td>
<td>24.589</td>
<td>0.793</td>
<td>0.757</td>
<td>1.86</td>
<td>-18.12</td>
<td>-18.11</td>
<td>-18.58</td>
</tr>
<tr>
<td>25.453</td>
<td>24.878</td>
<td>0.952</td>
<td>0.964</td>
<td>0.72</td>
<td>-17.77</td>
<td>-17.81</td>
<td>-18.06</td>
</tr>
<tr>
<td>25.911</td>
<td>24.994</td>
<td>0.921</td>
<td>0.890</td>
<td>0.96</td>
<td>-17.34</td>
<td>-17.59</td>
<td>-18.11</td>
</tr>
<tr>
<td>26.004</td>
<td>24.915</td>
<td>0.986</td>
<td>0.966</td>
<td>0.95</td>
<td>-17.38</td>
<td>-16.98</td>
<td>-17.30</td>
</tr>
<tr>
<td>26.803</td>
<td>25.232</td>
<td>1.044</td>
<td>1.400</td>
<td>0.78</td>
<td>-16.67</td>
<td>-18.17</td>
<td>-19.17</td>
</tr>
<tr>
<td>25.204</td>
<td>25.314</td>
<td>0.929</td>
<td>0.882</td>
<td>0.64</td>
<td>-18.05</td>
<td>-18.68</td>
<td>-19.63</td>
</tr>
<tr>
<td>25.357</td>
<td>24.735</td>
<td>0.901</td>
<td>0.875</td>
<td>1.69</td>
<td>-17.64</td>
<td>-17.48</td>
<td>-17.67</td>
</tr>
<tr>
<td>24.117</td>
<td>24.028</td>
<td>0.484</td>
<td>0.511</td>
<td>0.84</td>
<td>-16.64</td>
<td>-16.60</td>
<td>-16.83</td>
</tr>
<tr>
<td>26.108</td>
<td>25.342</td>
<td>0.763</td>
<td>1.400</td>
<td>1.07</td>
<td>-16.27</td>
<td>-16.39</td>
<td>-15.54</td>
</tr>
<tr>
<td>24.909</td>
<td>25.120</td>
<td>0.711</td>
<td>1.152</td>
<td>0.42</td>
<td>-17.09</td>
<td>-17.21</td>
<td>-17.85</td>
</tr>
<tr>
<td>24.474</td>
<td>24.681</td>
<td>1.044</td>
<td>1.096</td>
<td>0.69</td>
<td>-18.95</td>
<td>-18.95</td>
<td>-19.22</td>
</tr>
<tr>
<td>23.100</td>
<td>24.234</td>
<td>0.826</td>
<td>1.391</td>
<td>0.53</td>
<td>-19.61</td>
<td>-19.85</td>
<td>-20.28</td>
</tr>
<tr>
<td>22.009</td>
<td>22.633</td>
<td>0.340</td>
<td>0.323</td>
<td>2.88</td>
<td>-17.49</td>
<td>-17.64</td>
<td>-18.17</td>
</tr>
</tbody>
</table>
The goals of multivariate analysis:

Generalize univariate statistical methods
   Multivariate means, variances, and covariances
   Multivariate probability distributions

Reduce the number of variables
   Structural simplification
   Linear functions of variables (principal components)

Investigate the dependence between variables
   Canonical correlations

Statistical inference
   Confidence regions
   Multivariate regression
   Hypothesis testing

Classify or cluster “similar” objects
   Discriminant analysis
   Cluster analysis

Prediction
Organizing the data

$p$: The number of variables

$n$: The number of objects (cases) (the sample size)

$x_{ij}$: the $i^{th}$ observation on the $j^{th}$ variable

Data array or data matrix

<table>
<thead>
<tr>
<th>Objects</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>...</td>
<td>$p$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>...</td>
<td>$x_{1p}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>...</td>
<td>$x_{2p}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| ...     | ... | ... | ...| ... | ... | ...
| $n$     | $x_{n1}$ | $x_{n2}$ | ...| $x_{np}$ |    |    |
Data matrix

\[
X = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1p} \\
  x_{21} & x_{22} & \cdots & x_{2p} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{np}
\end{bmatrix}
\]

We write \(X\) as \(n\) row or as \(p\) column vectors

\[
X = \begin{bmatrix}
  x_1^T \\
  x_2^T \\
  \vdots \\
  x_n^T
\end{bmatrix} = [y_1, y_2, \ldots, y_p]
\]

Matrix methods are essential to multivariate analysis

We will need only small amounts of matrix methods, e.g.,

\(A^T\): The transpose of \(A\)

\(|A|\): The determinant of \(A\)

\((AB)^T = B^T A^T\)
**Descriptive Statistics**

The sample mean of the $j^{th}$ variable:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

The sample mean vector:

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

The sample variance of the $j^{th}$ variable:

$$s_{jj} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2$$

The sample covariance of variables $i$ and $j$:

$$s_{ij} = s_{ji} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

**Question:** Why do we divide by $(n - 1)$ rather than $n$?
The sample covariance matrix:

\[ S = \begin{bmatrix}
    s_{11} & s_{12} & \cdots & s_{1p} \\
    s_{21} & s_{22} & \cdots & s_{2p} \\
    \vdots & \vdots & \ddots & \vdots \\
    s_{p1} & s_{p2} & \cdots & s_{pp}
\end{bmatrix} \]

The sample correlation coefficient of variables \( i \) and \( j \):

\[ r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \]

Note that \( r_{ii} = 1 \) and \( r_{ij} = r_{ji} \)

The sample correlation matrix:

\[ R = \begin{bmatrix}
    1 & r_{12} & \cdots & r_{1p} \\
    r_{21} & 1 & \cdots & r_{2p} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{p1} & r_{p2} & \cdots & 1
\end{bmatrix} \]
S and R are symmetric

S and R are positive semidefinite: $v^T S v \geq 0$ for any vector $v$.

Equivalently,

\[ s_{11} \geq 0, \quad \begin{vmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{vmatrix} \geq 0, \quad \begin{vmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{vmatrix} \geq 0, \]

etc.

If S is singular so is R and conversely.

If $n \leq p$ then S and R will be singular:

\[ |S| = 0 \text{ and } |R| = 0 \]

Which practical astrophysicist would attempt a statistical analysis with 65 variables and a sample size smaller than 65?
$v^T S v > 0$ is the variance of $v^T X$

If $n > p$ then, generally (but not always), $S$ and $R$ are strictly positive definite:

Then $\text{Var}(v^T X) = v^T S v > 0$ for any non-zero vector $v$

Equivalently,

$$s_{11} > 0, \quad \begin{vmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{vmatrix} > 0,$$

etc.

However, if $n > p$ and $|S| = 0$ then for some $v$ $\text{Var}(v^T X) = 0$ implying $v^T X$ is a constant and there is a linear relationship between the components of $X$

In this case, we can eliminate the dependent variables: **dimension reduction**
The COMBO-17 data

Variables: Rmag, $\mu_{\text{max}}$, Mcz, MCzml, chi2red, UjMAG, BjMAG, VjMAG

$p = 8$ and $n = 3462$

The sample mean vector:

\[
R_{\text{mag}} \quad \mu_{\text{max}} \quad \text{Mc}z \quad \text{MCzml} \quad \text{chi}2\text{red} \quad U_{\text{MAG}} \quad B_{\text{MAG}} \quad V_{\text{MAG}}
\]
\[
23.939 \quad 24.182 \quad 0.729 \quad 0.770 \quad 1.167 \quad -17.866 \quad -17.749 \quad -18.113
\]

The sample covariance matrix:

\[
\begin{array}{cccccccc}
R_{\text{mag}} & \mu_{\text{max}} & \text{Mc}z & \text{MCzml} & \text{chi}2\text{red} & U_{\text{MAG}} & B_{\text{MAG}} & V_{\text{MAG}} \\
R_{\text{mag}} & 2.062 & 1.362 & 0.190 & 0.234 & 0.147 & 0.890 & 1.015 & 1.060 \\
\mu_{\text{max}} & 1.362 & 1.035 & 0.141 & 0.172 & 0.079 & 0.484 & 0.578 & 0.610 \\
\text{Mc}z & 0.190 & 0.141 & 0.102 & 0.105 & -0.004 & -0.438 & -0.425 & -0.428 \\
\text{MCzml} & 0.234 & 0.172 & 0.105 & 0.141 & -0.009 & -0.416 & -0.414 & -0.419 \\
\text{chi}2\text{red} & 0.147 & 0.079 & -0.004 & -0.009 & 0.466 & 0.201 & 0.204 & 0.221 \\
U_{\text{MAG}} & 0.890 & 0.484 & -0.438 & -0.416 & 0.201 & 3.863 & 3.890 & 3.946 \\
B_{\text{MAG}} & 1.015 & 0.578 & -0.425 & -0.414 & 0.204 & 3.890 & 4.500 & 4.219 \\
V_{\text{MAG}} & 1.060 & 0.610 & -0.428 & -0.419 & 0.221 & 3.946 & 4.219 & 4.375 \\
\end{array}
\]
Advice given by some for Correlation Matrix:

- Use no more than two significant digits.

- Starting with the physically most important variable, reorder variables by descending correlations.

- Suppress diagonal entries to ease visual clutter.

- Suppress zeros before the decimal point.

COMBO-17's correlation matrix

<table>
<thead>
<tr>
<th></th>
<th>Rmag</th>
<th>mumax</th>
<th>Mcz</th>
<th>MCzml</th>
<th>chi2red</th>
<th>UjMAG</th>
<th>BjMAG</th>
<th>VjMAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rmag</td>
<td>.9</td>
<td>.4</td>
<td>.4</td>
<td>.2</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td>mumax</td>
<td>.9</td>
<td>.4</td>
<td>.5</td>
<td>.1</td>
<td>.2</td>
<td>.3</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td>Mcz</td>
<td>.4</td>
<td>.4</td>
<td>.9</td>
<td>-.0</td>
<td>-.7</td>
<td>-.6</td>
<td>-.6</td>
<td></td>
</tr>
<tr>
<td>MCzml</td>
<td>.4</td>
<td>.5</td>
<td>.9</td>
<td>-.0</td>
<td>-.6</td>
<td>-.5</td>
<td>-.5</td>
<td></td>
</tr>
<tr>
<td>chi2red</td>
<td>.2</td>
<td>.1</td>
<td>-.0</td>
<td>-.0</td>
<td>.2</td>
<td>.1</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>UjMAG</td>
<td>.3</td>
<td>.2</td>
<td>-.7</td>
<td>-.6</td>
<td>.2</td>
<td>.9</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>BjMAG</td>
<td>.3</td>
<td>.3</td>
<td>-.6</td>
<td>-.5</td>
<td>.1</td>
<td>.9</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>VjMAG</td>
<td>.4</td>
<td>.3</td>
<td>-.6</td>
<td>-.5</td>
<td>.2</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>
**Reminder:** Correlations measure the strengths of linear relationships between variables *if* such relationships are valid

\{UjMAG, BjMAG, VjMAG\} are highly correlated; perhaps, two of them can be eliminated. Similar remarks apply to \{Rmag, mumax\} and \{Mcz, Mczml\}.

chi2red has small correlation with \{mumax, Mcz, Mczml\}; we would retain chi2red in the subsequent analysis.
Multivariate probability distributions

Find the \textit{probability} that a galaxy chosen \textit{at random} from the population of \textit{all} COMBO-17 type galaxies satisfies

\[ 4 \times \text{Rmag} + 3 \times \text{mumax} + |\text{Mcz-MCzml}| - \chi^{2}_{\text{red}} + (U_j\text{MAG}+B_j\text{MAG})^2 + V_j\text{MAG}^2 < 70? \]

\( X_1: \text{Rmag} \)

\( X_2: \text{mumax} \)

\( \ldots \)

\( X_7: \text{BjMAG} \)

\( X_8: \text{VjMAG} \)
We wish to make probability statements about random vectors

$p$-dimensional random vector:

$$
\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}
$$

where $X_1, \ldots, X_p$ are random variables

$\mathbf{X}$ is a *continuous random vector* if $X_1, \ldots, X_p$ all are continuous random variables

We shall concentrate on continuous random vectors
Each nice $X$ has a prob. density function $f$

Three important properties of the p.d.f.:

1. $f(x) \geq 0$ for all $x \in \mathbb{R}^p$

2. The total area below the graph of $f$ is 1:
   \[
   \int_{\mathbb{R}^p} f(x) \, dx = 1
   \]

3. For all $t_1, \ldots, t_p$,
   \[
   P(X_1 \leq t_1, \ldots, X_p \leq t_p) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_p} f(x) \, dx
   \]
Reminder: “Expected value,” an average over the entire population

The mean vector:

\[
\mu = \begin{bmatrix}
\mu_1 \\
\vdots \\
\mu_p
\end{bmatrix}
\]

where

\[
\mu_i = E(X_i) = \int_{\mathbb{R}^p} x_i f(x) \, dx
\]

is the mean of the \(i\)th component of \(X\)

The covariance between \(X_i\) and \(X_j\):

\[
\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j) = E(X_i X_j) - \mu_i \mu_j
\]

The variance of each \(X_i\):

\[
\sigma_{ii} = E(X_i - \mu_i)^2 = E(X_i^2) - \mu_i^2
\]
The covariance matrix of $X$:

$$
\Sigma = 
\begin{bmatrix}
\sigma_{11} & \cdots & \sigma_{1p} \\
\sigma_{21} & \cdots & \sigma_{2p} \\
\vdots & \ddots & \vdots \\
\sigma_{p1} & \cdots & \sigma_{pp}
\end{bmatrix}
$$

An easy result:

$$
\Sigma = E(X - \mu)(X - \mu)^T
$$

Also,

$$
\Sigma = E(XX^T) - \mu\mu^T
$$

To avoid pathological cases, we assume that $\Sigma$ is nonsingular
### Theory vs. Practice

### Population vs. Random Sample

<table>
<thead>
<tr>
<th>All galaxies of COMBO-17 type</th>
<th>A sample from the COMBO-17 data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random vector $X$</td>
<td>Random sample $x_1, \ldots, x_n$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population Mean $\mu = E(X)$</th>
<th>Sample mean $\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Popn. cov. matrix $\Sigma =$</th>
<th>Sample cov. matrix, $S = \frac{1}{n-1}$ $\times \sum (x_k - \bar{x})(x_k - \bar{x})^T$</th>
</tr>
</thead>
</table>

$$E(X - \mu)(X - \mu)^T$$

Laws of Large Numbers: In a technical sense, $\bar{x} \to \mu$ and $S \to \Sigma$ as $n \to \infty$
The Multivariate Normal Distribution

\( \mathbf{X} = [X_1, \ldots, X_p]^T \): A random vector whose possible values range over all of \( \mathbb{R}^p \)

\( \mathbf{X} \) has a multivariate normal distribution if has a probability density function of the form

\[
    f(\mathbf{x}) = \text{const.} \times \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]
\]

where

\[
    \text{const.} = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}}
\]

Standard notation: \( \mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma) \)

Special case, \( p = 1 \): Let \( \Sigma = \sigma^2 \); then

\[
    f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right]
\]
Special case, $\Sigma$ diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$$

$$|\Sigma| = \sigma_1^2 \sigma_2^2 \cdots \sigma_p^2$$

$$\Sigma^{-1} = \begin{bmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_p^{-2} \end{bmatrix}$$

$$f(x) = \prod_{j=1}^{p} \frac{1}{(2\pi \sigma_j^2)^{1/2}} \exp \left[ -\frac{1}{2} \left( \frac{x_j - \mu_j}{\sigma_j} \right)^2 \right]$$

Conclusion: $X_1, \ldots, X_p$ are mutually independent and normally distributed
Recall: $X \sim N_p(\mu, \Sigma)$ if its p.d.f. is of the form

$$f(x) = \text{const.} \times \exp \left[ -\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu) \right]$$

where

$$\text{const.} = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}}$$

Facts:

$$\mu = E(X),$$

$$\Sigma = \text{Cov}(X)$$

$$\int_{\mathbb{R}^p} f(x) \, dx = 1$$
If $A$ is a $k \times p$ matrix then

$$AX + b \sim N_k(A\mu + b, A\Sigma A^T)$$

Proof: Use Fourier transforms

Special cases:

$b = 0$ and $A = v^T$ where $v \neq 0$:

$$v^TX \sim N(v^T\mu, v^T\Sigma v)$$

Note: $v^T\Sigma v > 0$ since $\Sigma$ is positive definite

$v = [1, 0, \ldots, 0]^T$: $X_1 \sim N(\mu_1, \sigma_{11})$

Similar argument: Each $X_i \sim N(\mu_i, \sigma_{ii})$
Decompose $X$ into two subsets, $X = \begin{bmatrix} X_u \\ X_l \end{bmatrix}$

Similarly, decompose

$$\mu = \begin{bmatrix} \mu_u \\ \mu_l \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{uu} & \Sigma_{ul} \\ \Sigma_{lu} & \Sigma_{ll} \end{bmatrix}$$

Then

$$\mu_u = E(X_u), \quad \mu_l = E(X_l)$$

$$\Sigma_{uu} = \text{Cov}(X_u), \quad \Sigma_{ll} = \text{Cov}(X_l)$$

$$\Sigma_{ul} = \text{Cov}(X_u, X_l)$$

The marginal distribution of $X_u$:

$$X_u \sim N_u(\mu_u, \Sigma_{uu})$$

The conditional distribution of $X_u|X_l$:

$$X_u|X_l \sim N_u(\cdots, \cdots)$$
If $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$ then $v^T \mathbf{X}$ has a 1-D normal distribution for every vector $v \in \mathbb{R}^p$

Conversely, if $v^T \mathbf{X}$ has a 1-D normal distribution for every $v$ then $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$

Proof: Fourier transforms again

(The assumption that an $\mathbf{X}$ is normally distributed is very strong)

Let us use this result to construct an exploratory test of whether some COMBO-17 variables have a multivariate normal distribution

Choose several COMBO-17 variables, e.g.,

$R_{\text{mag}}$, $\mu_{\text{max}}$, $M_{\text{cz}}$, $M_{\text{Czml}}$, $\chi^2_{\text{red}}$, $U_{\text{J MAG}}$, $B_{\text{J MAG}}$, $V_{\text{J MAG}}$
Use R to generate a “random” vector \( \mathbf{v} = [v_1, v_2, \ldots, v_8]^T \)

For each galaxy, calculate

\[ v_1 \cdot R\text{mag} + v_2 \cdot \mu\text{max} + \cdots + v_8 \cdot V_j\text{MAG} \]

This produces 3,462 such numbers (\( v \)-scores)

Construct a Q-Q plot of all these \( v \)-scores against the standard normal distribution

Study the plot to see if normality seems plausible

Repeat the exercise with a new random \( \mathbf{v} \)

Repeat the exercise \( 10^3 \) times

Note: We need only those vectors for which

\[ v_1^2 + \cdots + v_8^2 = 1 \] (why?)
Mardia’s test for multivariate normality

If the data contain a substantial number of outliers then it goes against the hypothesis of multivariate normality

If one COMBO-17 variable is not normally distributed then the full set of variables does not have a multivariate normal distribution

In that case, we can try to transform the original variables to produce new variables which are normally distributed

Example: Box-Cox transformations, log transformations (a special case of Box-Cox)

For data sets arising from a multivariate normal distribution, we can perform accurate inference for the mean vector and covariance matrix
Variables (random vector): $\mathbf{X} \sim N_p(\mu, \Sigma)$

The parameters $\mu$ and $\Sigma$ are unknown

Data (measurements): $x_1, \ldots, x_n$

Problem: Estimate $\mu$ and $\Sigma$

$\bar{x}$ is an unbiased and consistent estimator of $\mu$

$\bar{x}$ is the MLE of $\mu$

The MLE of $\Sigma$ is $\frac{n-1}{n} S$; this is not unbiased

The sample covariance matrix, $S$, is an unbiased estimator of $\Sigma$

Since $S$ is close to being the MLE of $\Sigma$, we estimate $\Sigma$ using $S$
A confidence region for $\mu$

Naive method: Using only the data on the $i$th variable, construct a confidence interval for each $\mu_i$.

Use the collection of confidence intervals as a confidence region for $\mu$.

Good news: This can be done using elementary statistical methods.

Bad news: A collection of 95% confidence intervals, one for each $\mu_i$, does not result in a 95% confidence region for $\mu$.

Starting with individual intervals with lower confidence levels, we can achieve an overall 95% confidence level for the combined region.

Bonferroni inequalities: Some difficult math formulas are needed to accomplish that goal.
Worse news: The resulting confidence region for $\mu$ is a rectangle

This is not consonant with a density function of the form

$$f(x) = \text{const.} \times \exp \left[ -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right]$$

The contours of the graph of $f(x)$ are ellipsoids, so we should derive an ellipsoidal confidence region for $\mu$
Fact: Every positive definite symmetric matrix has a unique positive definite symmetric square root

$\Sigma^{-1/2}$: The p.d. square-root of $\Sigma^{-1}$

Recall (see p. 31): If $A$ is a $p \times p$ nonsingular matrix and $X \sim N_p(\mu, \Sigma)$ then

$$AX + b \sim N_p(A\mu + b, A\Sigma A^T)$$

Set $A = \Sigma^{-1/2}$, $b = -\Sigma^{-1/2}\mu$

Then $A\mu + b = 0$, $A\Sigma A^T = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_p$

$$\Sigma^{-1/2}(X - \mu) \sim N_p(0, I_p)$$

$I_p = \text{diag}(1, 1, \ldots, 1)$, a diagonal matrix
Methods of Multivariate Analysis

Reduce the number of variables
   Structural simplification
   Linear functions of variables (Principal Components)

Investigate the dependence between variables
   Canonical correlations

Statistical inference
   Estimation
   Confidence regions
   Hypothesis testing

Classify or cluster “similar” objects
   Discriminant analysis
   Cluster analysis

Predict
   Multiple Regression
   Multivariate regression
Principal Components Analysis (PCA)

COMBO-17: $p = 65$ (wow!)

Can we reduce the dimension of the problem?

$X$: A $p$-dimensional random vector

Covariance matrix: $\Sigma$

Solve for $\lambda$: $|\Sigma - \lambda I| = 0$

Solutions: $\lambda_1, \ldots, \lambda_p$, the eigenvalues of $\Sigma$

Assume, for simplicity, that $\lambda_1 > \cdots > \lambda_p$

Solve for $v$: $\Sigma v = \lambda_j v$, $j = 1, \ldots, p$

Solution: $v_1, \ldots, v_p$, the eigenvectors of $\Sigma$

Scale each eigenvector to make its length 1

$v_1, \ldots, v_p$ are orthogonal
The first PC: The linear combination $v^T X$ such that 
(i) $\text{Var}(v^T X)$ is maximal, and 
(ii) $v^T v = 1$

Maximize $\text{Var}(v^T X) = v^T \Sigma v$ subject to $v^T v = 1$

Lagrange multipliers

Solution: $v = v_1$, the first eigenvector of $\Sigma$

$v_1^T X$ is the first principal component

The second PC: The linear combination $v^T X$ such that 
(i) $\text{Var}(v^T X)$ is maximal, 
(ii) $v^T v = 1$, and 
(iii) $v^T X$ has zero correlation with the first PC
Maximize $\text{Var}(v^T X) = v^T \Sigma v$ with $v^T v = 1$ and 
$\text{Cov}(v^T X, v_1^T X) \equiv v^T \Sigma v_1 = 0$

Lagrange multipliers

Solution: $v = v_2$, the second eigenvector of $\Sigma$

The $k$th PC: The linear combination $v^T X$ such that

(i) $\text{Var}(v^T X)$ is maximal,
(ii) $v^T v = 1$, and
(iii) $v^T X$ has zero correlation with all prior PCs

Solution: $v = v_k$, the $k$th eigenvector of $\Sigma$

The PCs are random variables

Simple matrix algebra: $\text{Var}(v_k^T X) = \lambda_k$
$p$-dimensional data: $x_1, \ldots, x_n$

$S$: the sample covariance matrix

$\tilde{\lambda}_1 > \cdots > \tilde{\lambda}_p$: The eigenvalues of $S$

Remarkable result:

$$\tilde{\lambda}_1 + \cdots + \tilde{\lambda}_p = s_{11} + \cdots + s_{pp}$$

$\tilde{v}_1, \ldots, \tilde{v}_p$: The corresponding eigenvectors

$\tilde{v}_1^T x, \ldots, \tilde{v}_p^T x$: The sample PCs

$\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p$: The estimated variances of the PCs

Basic idea: Use the sample PCs instead of $X$ to analyze the data
The sample principal components:

\[
S = \begin{bmatrix}
4.31 & 1.68 & 1.80 & 2.16 & -.25 \\
1.68 & 1.77 & .59 & .18 & .17 \\
1.80 & .59 & .80 & 1.07 & -.16 \\
2.16 & .18 & 1.07 & 1.97 & -.36 \\
-.25 & .17 & -.16 & -.36 & .50
\end{bmatrix}
\]

The sample principal components:

\[
Y_1 = .8X_1 + .3X_2 + .3X_3 + .4X_4 - .1X_5 \\
Y_2 = -.1X_1 - .8X_2 + .1X_3 + .6X_4 - .3X_5 \\
\text{etc.}
\]

\[
\tilde{\lambda}_1 = 6.9, \tilde{\lambda}_2 = 1.8, \ldots; \tilde{\lambda}_1 + \cdots + \tilde{\lambda}_5 = 8.4
\]

\(X_1\): Rmag \\
\(X_2\): mumax \\
\text{etc.}

The PCs usually have no physical meaning, but they can provide insight into the data analysis.
\(\tilde{\lambda}_1 + \cdots + \tilde{\lambda}_p\): A measure of total variability of the data

\[
\frac{\tilde{\lambda}_k}{\tilde{\lambda}_1 + \cdots + \tilde{\lambda}_p}: \text{ The proportion of total variability of the data } \text{“explained” by the } k\text{th PC}
\]

How many PC’s should we calculate?

Stop when

\[
\frac{\tilde{\lambda}_1 + \cdots + \tilde{\lambda}_k}{\tilde{\lambda}_1 + \cdots + \tilde{\lambda}_p} \geq 0.9
\]

Scree plot: Plot the points \((1, \tilde{\lambda}_1), \ldots, (p, \tilde{\lambda}_p)\) and connect them by a straight line. Stop when the graph has flattened.

Other rule: Kaiser’s rule; rules based on tests of hypotheses, ...
Some feel that PC’s should be calculated from correlation matrices, not covariance matrices

Argument for correlation matrices: If the original data are rescaled then the PCs and the $\tilde{\lambda}_k$ all change

Argument against: If some components have significantly smaller means and variances than others then correlation-based PCs will give all components similarly-sized weights
**COMBO-17 data:**

Two classes of galaxies, redder and bluer, but with overlapping distributions

Dataset: galaxy brightnesses in 17 bands—detailed view of "red" and "blue" for each galaxy

The following figure of $M_B$ (BjMag) vs (280-B) (S280MAG-BjMag) for restricted range 0.7-0.9 of z (McZ) shows two cluster ("blue" below and "red" above), similar to the one in the website (also Wolf et al., 2004)
An exercise:

We investigate the relationship of these colors to the brightness variables by multivariate analysis.

From combo17 dataset collected the even-numbered columns (30, 32, ..., 54).

Normalized each to (say) the value in column 40 (W640FE) for each galaxy. These are called “colors”.

Removed variable W640FE from the dataset

We added to this dataset Bjmag ($M_B$). Also kept Mcz.

Modified “W” variables have been renamed with an “R” in the beginning.
mean locations in multidimensional parameter space for "dust-free old" (≡ "red") and "blue cloud" (≡ "blue") galaxies

red galaxies has a mean value of \((U - V) = 1.372\)

blue galaxies has a mean \((U - V) = 0.670\)—which are widely separated values

redshift \(z\) is a scientifically (very!) interesting variable denoting age of galaxy

We classify as "red" if \((U - V) > 0.355\) and as "blue" if \((U - V) \leq 0.355\)—color variable

This is the dataset. Data for the first few galaxies with the first few “R” readings:
<table>
<thead>
<tr>
<th>RW420FE</th>
<th>RW462FE</th>
<th>RW485FD</th>
<th>RW518FE</th>
<th>RW571FS</th>
<th>RW604FE</th>
<th>BJMAG</th>
<th>MCZ</th>
<th>U-V</th>
<th>COLOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.018</td>
<td>-0.006</td>
<td>0.000</td>
<td>-0.001</td>
<td>-0.004</td>
<td>-0.002</td>
<td>-17.540</td>
<td>0.832</td>
<td>0.090</td>
<td>1</td>
</tr>
<tr>
<td>-0.003</td>
<td>0.002</td>
<td>-0.000</td>
<td>-0.002</td>
<td>0.007</td>
<td>0.006</td>
<td>17.860</td>
<td>0.927</td>
<td>-0.080</td>
<td>1</td>
</tr>
<tr>
<td>-0.010</td>
<td>-0.003</td>
<td>0.002</td>
<td>-0.007</td>
<td>-0.000</td>
<td>0.000</td>
<td>-19.910</td>
<td>1.202</td>
<td>0.660</td>
<td>2</td>
</tr>
<tr>
<td>0.006</td>
<td>0.010</td>
<td>-0.005</td>
<td>-0.004</td>
<td>-0.005</td>
<td>0.003</td>
<td>-17.390</td>
<td>0.912</td>
<td>-0.160</td>
<td>1</td>
</tr>
<tr>
<td>0.002</td>
<td>0.005</td>
<td>0.002</td>
<td>0.010</td>
<td>0.004</td>
<td>0.007</td>
<td>-18.400</td>
<td>0.848</td>
<td>1.680</td>
<td>2</td>
</tr>
<tr>
<td>0.004</td>
<td>0.004</td>
<td>0.005</td>
<td>0.002</td>
<td>0.005</td>
<td>0.005</td>
<td>-18.110</td>
<td>0.882</td>
<td>-0.520</td>
<td>1</td>
</tr>
<tr>
<td>-0.004</td>
<td>-0.009</td>
<td>-0.008</td>
<td>-0.011</td>
<td>-0.008</td>
<td>-0.011</td>
<td>-16.060</td>
<td>0.896</td>
<td>-0.860</td>
<td>1</td>
</tr>
<tr>
<td>-0.002</td>
<td>-0.005</td>
<td>-0.006</td>
<td>-0.000</td>
<td>-0.004</td>
<td>0.002</td>
<td>-16.490</td>
<td>0.930</td>
<td>0.140</td>
<td>1</td>
</tr>
<tr>
<td>0.018</td>
<td>0.017</td>
<td>0.008</td>
<td>0.020</td>
<td>0.011</td>
<td>0.015</td>
<td>-17.680</td>
<td>0.774</td>
<td>0.200</td>
<td>1</td>
</tr>
<tr>
<td>0.006</td>
<td>0.007</td>
<td>0.001</td>
<td>-0.004</td>
<td>-0.004</td>
<td>-0.000</td>
<td>-11.150</td>
<td>0.062</td>
<td>-0.310</td>
<td>1</td>
</tr>
<tr>
<td>-0.009</td>
<td>-0.007</td>
<td>-0.010</td>
<td>-0.009</td>
<td>-0.009</td>
<td>-0.008</td>
<td>-16.990</td>
<td>0.874</td>
<td>-0.030</td>
<td>1</td>
</tr>
<tr>
<td>-0.032</td>
<td>-0.021</td>
<td>-0.018</td>
<td>-0.024</td>
<td>-0.019</td>
<td>-0.020</td>
<td>-13.990</td>
<td>0.173</td>
<td>0.650</td>
<td>2</td>
</tr>
<tr>
<td>-0.015</td>
<td>-0.009</td>
<td>-0.013</td>
<td>-0.006</td>
<td>-0.013</td>
<td>-0.014</td>
<td>-18.490</td>
<td>1.109</td>
<td>0.190</td>
<td>1</td>
</tr>
<tr>
<td>0.002</td>
<td>-0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.012</td>
<td>0.002</td>
<td>-10.300</td>
<td>0.143</td>
<td>0.870</td>
<td>2</td>
</tr>
<tr>
<td>-0.028</td>
<td>-0.023</td>
<td>-0.020</td>
<td>-0.020</td>
<td>-0.025</td>
<td>-0.017</td>
<td>-18.210</td>
<td>0.626</td>
<td>1.360</td>
<td>2</td>
</tr>
<tr>
<td>0.011</td>
<td>0.015</td>
<td>-0.002</td>
<td>-0.003</td>
<td>0.002</td>
<td>0.009</td>
<td>-20.140</td>
<td>1.185</td>
<td>-0.200</td>
<td>1</td>
</tr>
<tr>
<td>0.010</td>
<td>0.007</td>
<td>0.012</td>
<td>0.010</td>
<td>0.010</td>
<td>0.015</td>
<td>-16.130</td>
<td>0.921</td>
<td>-1.570</td>
<td>1</td>
</tr>
<tr>
<td>0.001</td>
<td>0.004</td>
<td>0.004</td>
<td>0.001</td>
<td>0.002</td>
<td>0.003</td>
<td>-19.420</td>
<td>0.832</td>
<td>-0.170</td>
<td>1</td>
</tr>
<tr>
<td>0.005</td>
<td>0.013</td>
<td>-0.002</td>
<td>0.008</td>
<td>0.007</td>
<td>0.007</td>
<td>-18.110</td>
<td>0.793</td>
<td>0.460</td>
<td>2</td>
</tr>
<tr>
<td>-0.007</td>
<td>-0.002</td>
<td>-0.009</td>
<td>-0.002</td>
<td>0.000</td>
<td>-0.008</td>
<td>-17.810</td>
<td>0.952</td>
<td>0.290</td>
<td>1</td>
</tr>
<tr>
<td>-0.004</td>
<td>-0.004</td>
<td>-0.007</td>
<td>-0.009</td>
<td>-0.007</td>
<td>-0.002</td>
<td>-17.590</td>
<td>0.921</td>
<td>0.770</td>
<td>2</td>
</tr>
<tr>
<td>-0.007</td>
<td>-0.008</td>
<td>-0.014</td>
<td>-0.004</td>
<td>-0.003</td>
<td>-0.002</td>
<td>-16.980</td>
<td>0.986</td>
<td>-0.080</td>
<td>1</td>
</tr>
<tr>
<td>0.008</td>
<td>-0.004</td>
<td>0.003</td>
<td>-0.001</td>
<td>-0.001</td>
<td>0.007</td>
<td>-18.170</td>
<td>1.044</td>
<td>2.500</td>
<td>2</td>
</tr>
<tr>
<td>-0.000</td>
<td>0.002</td>
<td>0.004</td>
<td>0.000</td>
<td>0.004</td>
<td>0.001</td>
<td>-18.680</td>
<td>0.929</td>
<td>1.580</td>
<td>2</td>
</tr>
<tr>
<td>0.002</td>
<td>0.003</td>
<td>0.008</td>
<td>-0.003</td>
<td>0.001</td>
<td>0.001</td>
<td>-17.480</td>
<td>0.901</td>
<td>0.030</td>
<td>1</td>
</tr>
<tr>
<td>0.020</td>
<td>0.013</td>
<td>0.009</td>
<td>0.009</td>
<td>0.018</td>
<td>0.026</td>
<td>-16.600</td>
<td>0.484</td>
<td>0.190</td>
<td>1</td>
</tr>
<tr>
<td>0.016</td>
<td>0.008</td>
<td>0.019</td>
<td>0.019</td>
<td>0.014</td>
<td>0.010</td>
<td>-16.390</td>
<td>0.763</td>
<td>-0.730</td>
<td>1</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
<td>0.006</td>
<td>0.004</td>
<td>0.003</td>
<td>0.002</td>
<td>-17.210</td>
<td>0.711</td>
<td>0.760</td>
<td>2</td>
</tr>
<tr>
<td>0.003</td>
<td>-0.001</td>
<td>-0.008</td>
<td>0.004</td>
<td>0.002</td>
<td>-0.001</td>
<td>-18.950</td>
<td>1.044</td>
<td>0.270</td>
<td>1</td>
</tr>
<tr>
<td>0.007</td>
<td>-0.007</td>
<td>-0.006</td>
<td>0.008</td>
<td>0.007</td>
<td>0.011</td>
<td>-19.850</td>
<td>0.826</td>
<td>0.670</td>
<td>2</td>
</tr>
<tr>
<td>-0.030</td>
<td>-0.013</td>
<td>-0.017</td>
<td>-0.001</td>
<td>0.021</td>
<td>0.025</td>
<td>-17.640</td>
<td>0.340</td>
<td>0.680</td>
<td>2</td>
</tr>
<tr>
<td>-0.058</td>
<td>-0.031</td>
<td>-0.037</td>
<td>-0.026</td>
<td>-0.015</td>
<td>-0.012</td>
<td>-17.600</td>
<td>0.365</td>
<td>0.390</td>
<td>2</td>
</tr>
<tr>
<td>0.004</td>
<td>0.006</td>
<td>0.008</td>
<td>0.013</td>
<td>0.018</td>
<td>0.021</td>
<td>-20.040</td>
<td>0.898</td>
<td>0.080</td>
<td>1</td>
</tr>
<tr>
<td>-0.005</td>
<td>-0.004</td>
<td>-0.006</td>
<td>-0.006</td>
<td>0.001</td>
<td>0.005</td>
<td>-19.540</td>
<td>0.878</td>
<td>0.290</td>
<td>1</td>
</tr>
<tr>
<td>-0.009</td>
<td>0.003</td>
<td>-0.009</td>
<td>-0.006</td>
<td>0.001</td>
<td>-0.007</td>
<td>-12.970</td>
<td>0.082</td>
<td>0.510</td>
<td>2</td>
</tr>
</tbody>
</table>
PCA of Combo17 data:

PCA of the 12 color variables RW420FE RW462FE .... RW856FD RW914FD
The scree plot suggests that two components are adequate.

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1 weight</th>
<th>PC2 weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW420FE</td>
<td>0.954</td>
<td>0.107</td>
</tr>
<tr>
<td>RW462FE</td>
<td>0.957</td>
<td>0.144</td>
</tr>
<tr>
<td>RW485FD</td>
<td>0.960</td>
<td>0.149</td>
</tr>
<tr>
<td>RW518FE</td>
<td>0.938</td>
<td>0.218</td>
</tr>
<tr>
<td>RW571FS</td>
<td>0.810</td>
<td>0.456</td>
</tr>
<tr>
<td>RW604FE</td>
<td>0.128</td>
<td>0.902</td>
</tr>
<tr>
<td>RW646FD</td>
<td>-0.897</td>
<td>0.326</td>
</tr>
<tr>
<td>RW696FE</td>
<td>-0.914</td>
<td>0.223</td>
</tr>
<tr>
<td>RW753FE</td>
<td>-0.913</td>
<td>0.252</td>
</tr>
<tr>
<td>RW815FS</td>
<td>-0.953</td>
<td>0.134</td>
</tr>
<tr>
<td>RW856FD</td>
<td>-0.970</td>
<td>0.110</td>
</tr>
<tr>
<td>RW914FD</td>
<td>-0.961</td>
<td>0.117</td>
</tr>
</tbody>
</table>

Variance explained 9.547 1.386
% Variance explained 79.555 11.553
Two components explain most of the variation (about 91%)

**Interpretation:**

**Principal Component 1:**
Weights are nearly the same in magnitude (except for RW604FE—insignificant)
RW4... and RW5... vs RW6... RW7.. RW8.. RW9..

**Principal Component 2:**
RW604E the main component
Rest are nearly equal and small
Two components complement each other
Plot of PC scores of galaxies can be used for classification
Will see this in the Cluster Analysis chapter
Classification Methods:

Two distinct types of classification problems—unsupervised and supervised

Unsupervised classification: Cluster Analysis: to find groups in the data objects objects within a group are similar

Example: what kinds of celestial objects are there—stars, planets, moons, asteroids, galaxies, comets, etc.

Multivariate (qualitative and quantitative) data on objects used

Characterize each type of object by these variables

Example: C. Wolf, M. E. Gray, and K. Meisenheimer (2008): Red-sequence galaxies with young stars and dust: The cluster Abell 901/902 seen with COMBO-17. Astronomy & Astrophysics classify galaxies into three classes with properties in the following table by cluster analysis
Mean properties of the three galaxy SED class samples.

<table>
<thead>
<tr>
<th>Property</th>
<th>Dust-free old</th>
<th>Dusty red-seq</th>
<th>Blue cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N_{\text{galaxy}})</td>
<td>294</td>
<td>168</td>
<td>333</td>
</tr>
<tr>
<td>(N_{\text{field contamination}})</td>
<td>6</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>(N_{\text{spectra}})</td>
<td>144</td>
<td>69</td>
<td>36</td>
</tr>
<tr>
<td>(z_{\text{spec}})</td>
<td>0.1646</td>
<td>0.1646</td>
<td>0.1658</td>
</tr>
<tr>
<td>(\sigma_{cz}/(1 + z)/(\text{km/s}))</td>
<td>939</td>
<td>1181</td>
<td>926</td>
</tr>
<tr>
<td>(z_{\text{spec, N}})</td>
<td>0.1625</td>
<td>0.1615</td>
<td>N/A</td>
</tr>
<tr>
<td>(z_{\text{spec, S}})</td>
<td>0.1679</td>
<td>0.1686</td>
<td>N/A</td>
</tr>
<tr>
<td>(\sigma_{cz,N}/(1 + z)/(\text{km/s}))</td>
<td>589</td>
<td>597</td>
<td>N/A</td>
</tr>
<tr>
<td>(\sigma_{cz,S}/(1 + z)/(\text{km/s}))</td>
<td>522</td>
<td>546</td>
<td>N/A</td>
</tr>
<tr>
<td>(\log(\Sigma_{10}(\text{Mpc}/h)^2))</td>
<td>2.188</td>
<td>1.991</td>
<td>1.999</td>
</tr>
<tr>
<td>(EW_e(OII)/\AA)</td>
<td>N/A</td>
<td>4.2 \pm 0.4</td>
<td>17.5 \pm 1.5</td>
</tr>
<tr>
<td>(EW_a(H\delta)/\AA)</td>
<td>2.3 \pm 0.5</td>
<td>2.6 \pm 0.5</td>
<td>4.5 \pm 1.0</td>
</tr>
<tr>
<td>age/Gyr</td>
<td>6.2</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>(E_B - V)</td>
<td>0.044</td>
<td>0.212</td>
<td>0.193</td>
</tr>
<tr>
<td>((U - V)_{\text{rest}})</td>
<td>1.372</td>
<td>1.293</td>
<td>0.670</td>
</tr>
<tr>
<td>(M_V,\text{rest})</td>
<td>-19.31</td>
<td>-19.18</td>
<td>-18.47</td>
</tr>
<tr>
<td>B - R</td>
<td>1.918</td>
<td>1.847</td>
<td>1.303</td>
</tr>
<tr>
<td>V - I</td>
<td>1.701</td>
<td>1.780</td>
<td>1.290</td>
</tr>
<tr>
<td>R - I</td>
<td>0.870</td>
<td>0.920</td>
<td>0.680</td>
</tr>
<tr>
<td>U - 420</td>
<td>0.033</td>
<td>-0.079</td>
<td>-0.377</td>
</tr>
<tr>
<td>420 - 464</td>
<td>0.537</td>
<td>0.602</td>
<td>0.560</td>
</tr>
<tr>
<td>464 - 518</td>
<td>0.954</td>
<td>0.827</td>
<td>0.490</td>
</tr>
<tr>
<td>604 - 646</td>
<td>0.356</td>
<td>0.339</td>
<td>0.238</td>
</tr>
<tr>
<td>753 - 815</td>
<td>0.261</td>
<td>0.274</td>
<td>0.224</td>
</tr>
</tbody>
</table>
Supervised Learning or Discriminant Analysis

Know that there are these three types of galaxies

Have **Training Samples** where an expert (supervisor) classifies units in the sample

Multivariate observations on the sample units available

A new object is seen on which multivariate observations made

Problem: Classify it in one or other of the groups

In discriminant Analysis we develop a formula for such classification
Formula arrived at by performing discriminant analysis of training data

Some assumptions are often made

Multivariate normality in each group with a common covariance matrix

Find a classification rule that minimizes misclassification

This leads to **Linear Discriminant Function**, a linear combination of observed variables
Discriminant Analysis Example

Use “R” data to develop a formula for classification into color 1 or 2

The linear discriminant function is

$$0.345 + \text{RW420FE} \times 14.277 - \text{RW462FE} \times 0.844 - \text{RW485FD} \times 36.890 + \text{RW518FE} \times 6.541 + \text{RW571FS} \times 2.249 + \text{RW604FE} \times 25.670 + \text{RW646FD} \times 18.331 + \text{RW696FE} \times 15.123 - \text{RW753FE} \times 29.072 - \text{RW815FS} \times 16.970 - \text{RW856FD} \times 16.467 + \text{RW914FD} \times 2.024$$

If this value is $> 0$ we classify a galaxy as 1 (red); else 2 (blue)

Using the formula on the training sample, we get an idea of the performance of the classification rule as follows:
<table>
<thead>
<tr>
<th>Actual</th>
<th>Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>1</td>
<td>2,111</td>
</tr>
<tr>
<td>2</td>
<td>1,020</td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>Total</td>
<td>3,131</td>
</tr>
</tbody>
</table>

This is not a very good classification rule—the chosen variables do not provide adequate separation between blue and red
Multiple Regression

If a supervisor had used the value of $U - V$ to classify the galaxies into red and blue, and if values of $U - V$ are indeed available, then why not use them rather than the red-blue classification?

$U - V$ data rather than color data in training sample

Leads to Multiple Regression Analysis

Develop a formula for prediction of $U - V$ in a new galaxy from "$R" values.

Results of such a multiple (linear) regression analysis:

Multiple correlation: a measure of how good the regression is: 0.344

Not very good—much as in Discriminant Analysis
Table below shows which "R" variables are useful for prediction of $U - V$: those with small $p$-values.

Regression Coefficients and their significance

<table>
<thead>
<tr>
<th>Effect</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>0.175</td>
<td>0.012</td>
<td>15.010</td>
<td>0.000</td>
</tr>
<tr>
<td>RW420FE</td>
<td>-1.624</td>
<td>0.839</td>
<td>-1.936</td>
<td>0.053</td>
</tr>
<tr>
<td>RW462FE</td>
<td>0.895</td>
<td>1.371</td>
<td>0.653</td>
<td>0.514</td>
</tr>
<tr>
<td>RW485FD</td>
<td>5.072</td>
<td>1.664</td>
<td>3.049</td>
<td>0.002</td>
</tr>
<tr>
<td>RW518FE</td>
<td>-1.921</td>
<td>1.199</td>
<td>-1.602</td>
<td>0.109</td>
</tr>
<tr>
<td>RW571FS</td>
<td>-1.126</td>
<td>1.178</td>
<td>-0.956</td>
<td>0.339</td>
</tr>
<tr>
<td>RW604FE</td>
<td>-4.636</td>
<td>1.456</td>
<td>-3.184</td>
<td>0.001</td>
</tr>
<tr>
<td>RW646FD</td>
<td>-2.345</td>
<td>1.340</td>
<td>-1.750</td>
<td>0.080</td>
</tr>
<tr>
<td>RW696FE</td>
<td>-2.729</td>
<td>0.917</td>
<td>-2.977</td>
<td>0.003</td>
</tr>
<tr>
<td>RW753FE</td>
<td>3.943</td>
<td>1.020</td>
<td>3.866</td>
<td>0.000</td>
</tr>
<tr>
<td>RW815FS</td>
<td>3.394</td>
<td>0.902</td>
<td>3.761</td>
<td>0.000</td>
</tr>
<tr>
<td>RW856FD</td>
<td>3.059</td>
<td>0.961</td>
<td>3.182</td>
<td>0.001</td>
</tr>
<tr>
<td>RW914FD</td>
<td>0.036</td>
<td>0.740</td>
<td>0.049</td>
<td>0.961</td>
</tr>
</tbody>
</table>

54