Lecture 22: Multivariate analysis and principal component analysis

• So far, we have been considering hypothesis testing and data modeling when a given quantity (e.g. X-ray photon counts) is observed as a function of an independent variable (e.g. photon energy)

• Another important task involves the analysis of a sample of objects for which a set of quantities has been measured. None is singled out as the independent variable.

Example: SAT-R, SAT-V, GPA
Multivariate data

• Suppose we have \( p \) quantities that have been measured for each of \( N \) objects in a sample:

• For each object, we can plot a vector, \( \mathbf{y} \), in \( p \)-dimensional space to represent the values of the \( p \) measured quantities: the \( k \)th component of \( \mathbf{y} \) represents the value of the \( k \)th measured quantity.

• The entire sample is represented by a set of \( N \) such vectors, \( \mathbf{y}^{(i)} \) for \( i = 1, N \)
Multivariate data

- We can plot the data on a $p$-dimensional scatter plot

(Here, $p = 3$, and $N = 8$)
Multivariate data

• The sample has a mean, \( \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)} \)

such that \( \bar{y}_k \) is the mean of the kth measurable quantity

For each measured quantity, we can define the variance,

\[
Var(y_k) = \frac{1}{N-1} \sum_{i=1}^{N} (y_k^{(i)} - \bar{y}_k)^2
\]
Covariance matrix

• We can also examine correlations between the quantities by defining the covariance matrix

\[ S_{kl} = \frac{1}{N-1} \sum_{i=1}^{N} (y_k^{(i)} - \bar{y}_k)(y_l^{(i)} - \bar{y}_l) \]

• The covariance matrix is clearly a p x p symmetric matrix

• The diagonal elements are just the variances: \( S_{kk} = \text{Var}(y_k) \)

• The off-diagonal elements represent correlations (because \( S_{km} \) is zero if \( y_k \) and \( y_m \) are uncorrelated)
Correlation matrix

- Elements of the covariance matrix are clearly not scale-invariant or - in general - dimensionless.

- It is often convenient to renormalize to obtain the correlation matrix

\[
R_{kl} = \frac{S_{kl}}{\sqrt{S_{kk}S_{ll}}}
\]

which is dimensionless, has diagonal elements equal to unity and off-diagonal elements in the range \(-1 \leq R_{kl} \leq +1\). These are called the linear correlation coefficients.
Principal component analysis

• Principal component analysis is a powerful tool for analysing multivariate data.

• In principal component analysis, we look for linear combinations of measured variables that are *uncorrelated* with each other

\[
z_1 = a_{11}y_1 + a_{12}y_2 + \ldots + a_{1p}y_p
\]
\[
z_2 = a_{21}y_1 + a_{22}y_2 + \ldots + a_{2p}y_p \quad \text{etc}...
\]

which we write \( z = A y \)

where \( A \) is an orthonormal matrix representing a rotation
Principal component analysis: graphical interpretation

The matrix $A$ rotates the axes to align them with the major and minor axes of the ellipsoid.

$y_2 = \text{GPA}$

$y_1 = \text{SAT (total)}$
Principal component analysis

• The covariance matrix for $z = A y$ is now diagonal:

$$S'_{mn}(z) = \text{Var}(z_m) \text{ for } m = n$$

$$= 0 \text{ otherwise}$$

How does this relate to the covariance matrix, $S$, for $y$?
Principal component analysis

\[
S'_{mn} = \frac{1}{N-1} \sum_{i=1}^{N} (A_{mk} y_k^{(i)} - A_{mk} \bar{y}_k)(A_{nl} y_l^{(i)} - A_{nl} \bar{y}_l)
\]

\[
= A_{nl} A_{mk} S_{kl}
\]

Using summation convention (sum over repeated indices \(i, m, k, n\))

In other words, \(S' = A S A^T = A S A^{-1}\)

\[\Rightarrow S = A^{-1} S' A\]

This is just the eigenvalue decomposition of \(S\): \(S = U D U^{-1}\), where \(D = S' = \text{diag}(\lambda_k)\) and the columns of \(U = A^T\) are the eigenvectors
Principal component analysis

So obtaining the principal components is entirely equivalent to finding the eigenvectors of the covariance matrix: each eigenvector gives us one principal component.

The corresponding eigenvalue tells us the variance of that principal component (PC), and the PCs with the largest variances are the most important.

A key application of PCA is to reduce the dimensionality of the problem in the case where the eigenvalues cover a wide dynamic range.

=> We need only consider the first few PC instead of all p PC, greatly simplifying the description of the correlations.
Scree graphs

We number the principal components in order of decreasing eigenvalue, and plot the eigenvalue as a function of its rank.

All information contained in first 3 PC

Dominated by noise
Preconditioning for scale invariance

We noted previously that the covariance matrix is not scale invariant: result of PCA can depend on the units used.

Often makes sense to precondition the variables by dividing each variable by the square root of its variance: effectively amounts to working with the correlation matrix in place of the covariance matrix.
Example 1: properties of winged aphids (Jeffers 1967)

Jeffers analysed a sample of 40 winged aphids for which 19 variables had been measured

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LENGTH</td>
<td>body length</td>
</tr>
<tr>
<td>WIDTH</td>
<td>body width</td>
</tr>
<tr>
<td>FORWING</td>
<td>forewing length</td>
</tr>
<tr>
<td>HINWING</td>
<td>hind-wing length</td>
</tr>
<tr>
<td>SPIRAC</td>
<td>number of spiracles</td>
</tr>
<tr>
<td>ANTSEG 1</td>
<td>length of antennal segment I</td>
</tr>
<tr>
<td>ANTSEG 2</td>
<td>length of antennal segment II</td>
</tr>
<tr>
<td>ANTSEG 3</td>
<td>length of antennal segment III</td>
</tr>
<tr>
<td>ANTSEG 4</td>
<td>length of antennal segment IV</td>
</tr>
<tr>
<td>ANTSEG 5</td>
<td>length of antennal segment V</td>
</tr>
<tr>
<td>ANTPIN</td>
<td>number of antennal spines</td>
</tr>
<tr>
<td>TARSUS 3</td>
<td>leg length, tarsus III</td>
</tr>
<tr>
<td>TIBIA 3</td>
<td>leg length, tibia III</td>
</tr>
<tr>
<td>FEMUR 3</td>
<td>leg length, femur III</td>
</tr>
<tr>
<td>ROSTRUM</td>
<td>rostrum</td>
</tr>
<tr>
<td>OVIPOS</td>
<td>ovipositor</td>
</tr>
<tr>
<td>OVSPIN</td>
<td>number of ovipositor spines</td>
</tr>
<tr>
<td>FOLD</td>
<td>anal fold</td>
</tr>
<tr>
<td>HOOKS</td>
<td>number of hind-wing hooks</td>
</tr>
</tbody>
</table>

From Renscher's *Methods of Multivariate analysis*
Correlation matrix for aphid properties

From Renscher's *Methods of Multivariate analysis*
# Eigenvalues of correlation matrix

Table 12.3  Eigenvalues of the Correlation Matrix of the Winged Aphid Data

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
<th>Percent of Variance</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.861</td>
<td>73.0</td>
<td>73.0</td>
</tr>
<tr>
<td>2</td>
<td>2.370</td>
<td>12.5</td>
<td>85.4</td>
</tr>
<tr>
<td>3</td>
<td>0.748</td>
<td>3.9</td>
<td>89.4</td>
</tr>
<tr>
<td>4</td>
<td>0.502</td>
<td>2.6</td>
<td>92.0</td>
</tr>
<tr>
<td>5</td>
<td>0.278</td>
<td>1.4</td>
<td>93.5</td>
</tr>
<tr>
<td>6</td>
<td>0.266</td>
<td>1.4</td>
<td>94.9</td>
</tr>
<tr>
<td>7</td>
<td>0.193</td>
<td>1.0</td>
<td>95.9</td>
</tr>
<tr>
<td>8</td>
<td>0.157</td>
<td>0.8</td>
<td>96.7</td>
</tr>
<tr>
<td>9</td>
<td>0.140</td>
<td>0.7</td>
<td>97.4</td>
</tr>
<tr>
<td>10</td>
<td>0.123</td>
<td>0.6</td>
<td>98.1</td>
</tr>
<tr>
<td>11</td>
<td>0.092</td>
<td>0.4</td>
<td>98.6</td>
</tr>
<tr>
<td>12</td>
<td>0.074</td>
<td>0.4</td>
<td>99.0</td>
</tr>
<tr>
<td>13</td>
<td>0.060</td>
<td>0.3</td>
<td>99.3</td>
</tr>
<tr>
<td>14</td>
<td>0.042</td>
<td>0.2</td>
<td>99.5</td>
</tr>
<tr>
<td>15</td>
<td>0.036</td>
<td>0.2</td>
<td>99.7</td>
</tr>
<tr>
<td>16</td>
<td>0.024</td>
<td>0.1</td>
<td>99.8</td>
</tr>
<tr>
<td>17</td>
<td>0.020</td>
<td>0.1</td>
<td>99.9</td>
</tr>
<tr>
<td>18</td>
<td>0.011</td>
<td>0.1</td>
<td>100.0</td>
</tr>
<tr>
<td>19</td>
<td>0.003</td>
<td>0.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

From Renscher's *Methods of Multivariate analysis*
Eigenvectors of correlation matrix

Table 12.4  Eigenvectors for the First Four Components of the Winged Aphid Data

| Variable   | Eigenvectors | | | |
|------------|--------------|---|---|---|---|
|            | 1       | 2    | 3    | 4    |
| LENGTH     | .96     | -.06 | .03  | -.12 |
| WIDTH      | .98     | -.12 | .01  | -.16 |
| FORWING    | .99     | -.06 | -.06 | -.11 |
| HINWING    | .98     | -.16 | .03  | -.00 |
| SPIRAC     | .61     | .74  | -.20 | 1.00 |
| ANTSEG 1   | .91     | .33  | .04  | .02  |
| ANTSEG 2   | .96     | .30  | .00  | -.04 |
| ANTSEG 3   | .88     | -.43 | .06  | -.18 |
| ANTSEG 4   | .90     | -.08 | .18  | -.01 |
| ANTSEG 5   | .94     | .05  | .11  | .03  |
| ANTSPIN    | -.49    | .37  | 1.00 | .27  |
| TARSUS 3   | .99     | -.02 | .03  | -.29 |
| TIBIA 3    | 1.00    | -.05 | .09  | -.31 |
| FEMUR 3    | .99     | -.12 | .12  | -.31 |
| ROSTRUM    | .96     | .02  | .08  | -.06 |
| OVIPOS     | .76     | .73  | -.03 | -.09 |
| OVSPIN     | .41     | 1.00 | -.16 | -.06 |
| FOLD       | -.71    | .64  | .04  | -.80 |
| HOOKS      | .76     | -.52 | .06  | .72  |

From Renscher's *Methods of Multivariate analysis*
Data points in the PC1 – PC2 plane

These clusters apparently represent 4 different species

Would be very hard to define these in terms of the original 19 parameters
Example 2: comparing spectral line maps in astronomy

- Spitzer Space Telescope provides spectroscopy in the 5.2 – 37 micron spectral range
Spectral lines can be mapped and show a variety of spatial distributions.
Principal component analysis

- Use PCA to write each map as a linear combination of orthogonal maps: $M_1, M_2, M_3 \ldots$ where $M_1$ accounts for most of the information, then $M_2$ etc. First two maps contain most of the information: 4th and higher components are noise dominated.
PCA allows the various transitions to be grouped

- Create an “h-plot”, in which we show the coefficients $A_{mk}$ for the first three maps ($m=1,2,3$)
PCA allows the various transitions to be grouped

Five distinct groups

1) $H_2 S(0)$
   Cool neutral gas
2) Other $H_2$ lines and $Si$
   Warm neutral gas
3) Fell, PII, SII
   Ionized gas
4) NeIII, NeII
   Highly ionized gas
5) SiIII
   Highly-ionized gas at low density
PCA allows the various transitions to be grouped

**Note:** cosine of angle between vectors represents correlation coefficient

\[
R_{k\ell} = A_{mk} A_{n\ell} R'_{n\ell}
\]

\[
= A_{mk} A_{n\ell} \delta_{mn}
\]

\[
= A_{mk} A_{m\ell}
\]

\[
= A_{1k} A_{1\ell} + A_{2k} A_{2\ell} = \cos \theta
\]

(if first two components dominate)