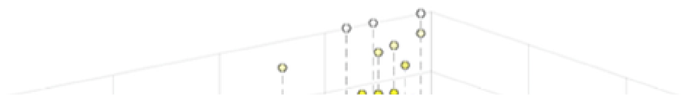
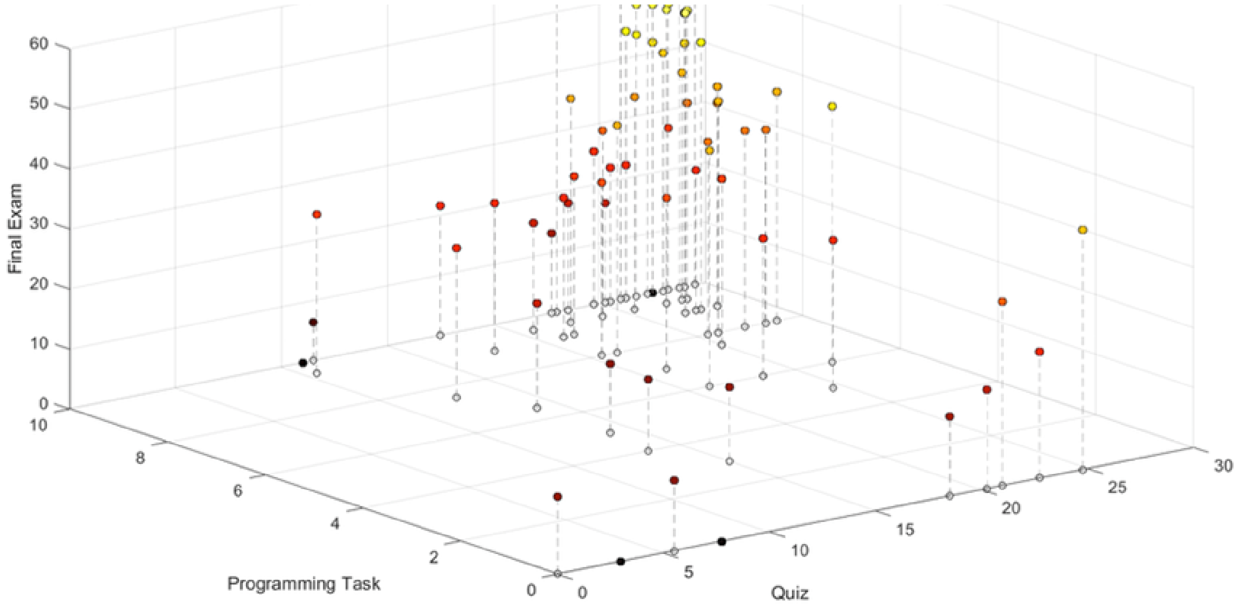
Big data: Mathematical modelling Principal Component Analysis

Background

Principal component analysis (PCA) can be used to analyse multivariate data that consists of lists of measurements made on a collection of objects or individuals. For example, consider an experiment for which $m = 3$ key observations (being O_1, O_2, O_3) are collected for $N = 73$ samples. In this case, we would store this data in what is refered to as a *matrix of observations*: X having 3 rows and 73 columns. We write $X \in \mathbb{R}^{m \times N}$ to reflect the shape of this matrix. One would say that the observed data is three-dimensional. One way to visualise this data is via the following scatter plot in three space $\mathbb{R}^3$.





The following MATLAB code demonstrates a simple PCA technique. Run the code in MATLAB and read the explanations below for further details.

```
% Load the required data set
load('student_data.mat');

% Extract the first 3 rows of interest
Xn = X(1:3,:);

% Visualise the raw data
scatter3(Xn(1,:)',Xn(2,:)',0*Xn(3,:)', 'MarkerFaceColor',[0.9 0.9 0.9],...
    'MarkerEdgeColor','k');
hold on;
scatter3(Xn(1,:)',Xn(2,:)',Xn(3,:)', 50 ,Xn(3,:)', 'filled',...
    'MarkerEdgeColor','k'); colormap('hot');
for i = 1:73
    plot3([Xn(1,i) Xn(1,i)],[Xn(2,i) Xn(2,i)], [0 Xn(3,i)], '--',...
        'Color', '[0.5 0.5 0.5]');
end
```

Denote the j^{th} column of X as $\mathbf{X}_{*j} = (O_{1j}, O_{2j}, O_{3j})^T$, for $j = 1, \ldots, N$ then $X = (\mathbf{X}_{*1}\,\mathbf{X}_{*2}\,\cdots\mathbf{X}_{*N})$. $\mathbf{X}_{*j}$ is often referred to as the *observation vector*. The symbol ^{T} represents taking the transpose of the given vector.

Our first tool used for data analysis is through what is termed a *correlation* matrix $C \in \mathbb{R}^{m\times m}$. To form C we proceed by first generating the sample mean ��et the vector containing all 1s be denoted by $\mathbf{e} = (1, 1, \ldots, 1)^T \in \mathbb{R}^{N\times 1}$ and compute �� $= \frac{1}{N}X\mathbf{e} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{X}_{*j}$Next, we subtract the sample mean

from each of the observation vectors to obtain a new matrix $\widehat{X}$ that is said to be in *mean-deviation* form, namely:

$$\widehat{X} = \left(\widehat{\mathbf{X}}_{*1}\,\widehat{\mathbf{X}}_{*2}\cdots\widehat{\mathbf{X}}_{*N}\right),$$

where $\widehat{\mathbf{X}}_{*j} = \mathbf{X}_{*j} - \boldsymbol{\mu}, j = 1, \ldots, N$. An interesting property for the mean-deviation form is that $\widehat{X}\,\mathbf{e} = \mathbf{0}_{m\times 1}$.

If we now normalise the rows of $\widehat{X}$ by scaling each of the i rows by their lengths, denoted here as $\frac{\widehat{\mathbf{X}}_{i*}}{\|\widehat{\mathbf{X}}_{i*}\|}$, and define $\widehat{X}_s$ as this scaled matrix, then the *correlation matrix* is given by

$$C = \widehat{X}_s \widehat{X}_s^T.$$

```
% Calculate how many observations there are
N = size(Xn, 2);
% Compute the average value of each variable
mu = sum(Xn,2)./N;

% Replicate the average vector into a matrix
mu_mat = repmat(mu,1,73);

% Subtract the average away from every column
Xhat = Xn - mu_mat;

% Preallocate the scaled matrix
Xhats = zeros(size(Xhat));

% Populate the scaled matrix by row
for i=1:3
    Xhats(i,:) = Xhat(i,:) / norm(Xhat(i,:));
end

% Compute the correlation matrix
C = Xhats * Xhats';

% Show the correlation matrix
C
```

Output

```
C =

    1.0000    0.5935    0.6964
    0.5935    1.0000    0.4665
    0.6964    0.4665    1.0000
```

Inspecting the correlation matrix, we can see that the first and third variables are most highly correlated, followed by variables 1 and 2, then 2 and 3.

For the discussion that follows, define vector $\mathbf{X} =(x_1, x_2, \ldots, x_m)^T$ as a vector that varies over the set of observation vectors $\mathbf{X}_{*j}, j = 1, \ldots, N$ so that x_1 represents a scalar (variable) that varies over the first coordinate of these vectors, x_2 a scalar that varies over the second coordinate, and so on. The $(i,j)^{th}$ entry of C (denoted $c_{i,j}$) represents the correlation between the variables x_i and x_j. An important interpretation is that two sets of variables may be compared by computing the cosine of the angle between the corresponding row vectors of $\widehat{X}$, which is precisely what the entries $c_{i,j}$ represent. A value near 1 in magnitude indicates that the two sets of variables are highly correlated, while a near 0 value indicates they are uncorrelated. We also note that if $c_{i,j} > 0$ the variables are said to be a positively correlated and if $c_{i,j} < 0$ they are negatively correlated.

Another important statistical tool used in data analysis is through the *sample covariance matrix* $S \in \mathbb{R}^{m \times m}$, which is very closely linked to the correlation matrix C defined above:

$$S = \frac{1}{N-1} \widehat{X} \widehat{X}^T.$$

Note that the matrix S is symmetric because $S^T = \frac{1}{N-1}\left(\widehat{X}\widehat{X}^T\right)^T = \frac{1}{N-1}(\widehat{X}^T)^T \widehat{X}^T = S$.

```
% Compute the sample covariance matrix
% Compute the sample covariance matrix
S = (1 / (N - 1)) * (Xhat * Xhat');

% Check that S is symmetric; the result should be a zero matrix
S' - S
```

Output

```
ans =

     0     0     0
     0     0     0
     0     0     0
```

A good way to obtain a better understanding of the meaning of the entries in S is to carry out partitioned matrix multiplication to obtain:

$$S = \frac{1}{N-1}\sum_{j=1}^{N}\left(\mathbf{X}_{*j} - \boldsymbol{\mu}\right)\left(\mathbf{X}_{*j} - \boldsymbol{\mu}\right)^T,$$

which is a representative form of a sample variance. The entries along the diagonal of S (denoted s_{ii}) are called the variances of the x_i, which measure the spread of the values x_i. A larger value implies a larger spread. The *total variance* of the data is the sum of the diagonal values of S, also known as the trace of S, ie, *Total Variance* = $\mathrm{tr}(S)$.

```
% Compute the trace of S, ie the total variance
trace(S)
```

Output

```
ans =

  309.7464
```

The off-diagonal entries s_{ij} in S represent the *covariance* of x_i and x_j. When these entries in S are zero, the corresponding variables are said to be uncorrelated. Note that the analysis of multivariate data is greatly simplified when most of the variables are uncorrelated, which is the case when S is a diagonal or near diagonal matrix.

We now come to principal component analysis (PCA). The primary objective of PCA is to determine an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ such that a clever change of variable $\widehat{X} = Q\widehat{Y}$ with $\widehat{Y} = (y_1, y_2, \ldots, y_m)^T$ renders the desirable property that the new variables y_i are uncorrelated and arranged in order of decreasing variance. A very important property of orthogonal matrices is that their inverse is simply the transpose of the matrix, ie, $QQ^T = Q^TQ = I_m$, where I_m is the $m \times m$ identity matrix.

Recall for the mean-deviation form discussed above that $\widehat{X}\,\mathbf{e} = \mathbf{0}_{m\times 1}$, we now show that the same property holds for $\widehat{Y}$. Consider $\widehat{Y}\mathbf{e} = Q^T\widehat{X}\,\mathbf{e} = Q^T\mathbf{0}_{m\times 1} = \mathbf{0}_{m\times 1}$. This result guarantees that the columns of $\widehat{Y}$ are also in mean-deviation form. Furthermore, the change of variable does not change the total variance of the data in the sense that $\mathrm{tr}(Q^TSQ) = \mathrm{tr}(Q^TQS) = \mathrm{tr}(S)$ using well-known properties of the trace.

Now, substituting our change of variable gives:

$$S = \frac{1}{N-1}(Q\widehat{Y})(Q\widehat{Y})^T,$$

or upon rearranging we have

or upon rearranging we have

$$Q^T S Q = \frac{1}{N-1}\widehat{Y}\widehat{Y}^T.$$

Clearly the best choice of Q is the one that makes $Q^T SQ$ a diagonal matrix. Before we answer how to find such a Q, we need to consider the eigenvalue problem.

Why does this work?

We now summarise some key facts for a symmetric matrix $S \in \mathbb{R}^{m\times m}$:

1. S has m real eigenvalues (taking into account their algebraic multiplicities), ie, if there are r distinct eigenvalues each having algebraic multiplicity $\text{alg mult}_S(\lambda_k)= a_k, k = 1, \ldots, r$ then $\sum_{k=1}^r a_k = m$.

2. The algebraic and geometric multiplicities for each distinct eigenvalue are equal, ie, $\text{alg mult}_S(\lambda)= \text{geo mult}_S(\lambda), \lambda \in \sigma(S)$.

3. Eigenvectors corresponding to different eigenvalues are orthogonal.

4. S is orthogonally diagonalisable; ie, let the set of orthonormal eigenvectors of S be given by $\mathbf{q}_k, \, k = 1, \ldots, m$ and set these as the columns of Q, then Q is orthogonal and $Q^T S Q = D$, where D is a diagonal matrix.

The steps to generating the orthonormal set of m eigenvectors $\mathbf{q}_k$ proceeds as follows. First, find a basis for each of the r eigenspaces of S. Next, apply the Gram-Schmidt process to each of these bases to obtain an orthonormal basis for each eigenspace. Form the matrix Q having these vectors as its columns.

Another property of S is that it is positive semi-definite because $\forall \mathbf{z} \in \mathbb{R}^m \backslash \{\mathbf{0}_{m\times 1}\}$, $\mathbf{z}^T S\mathbf{z} = \frac{1}{N-1}\mathbf{z}^T \widehat{X}\widehat{X}^T\mathbf{z} = \frac{1}{N-1}\mathbf{w}^T\mathbf{w} = \frac{1}{N-1}\|\mathbf{w}\|^2 \geq 0$, where $\mathbf{w} = \widehat{X}^T\mathbf{z}$. Equality occurs when $\mathbf{w} = \widehat{X}^T\mathbf{z} = \mathbf{0}_{N\times 1}$, which may arise in the case that $\text{rank}(\widehat{X}^T) < m$. What this result essentially tells us is that the eigenvalues of S are nonnegative. We will understand why this is the case shortly.

Let the set of eigenpairs of S be given by $\{(\lambda_k, \mathbf{q}_k)\}_{k=1}^m$ and define the columns of Q to be the normalised eigenvectors $\widehat{\mathbf{q}}_k = \frac{\mathbf{q}_k}{\|\mathbf{q}_k\|}$ (also known as the principal components of the data), then Q is orthogonal as a result of the properties listed above. Furthermore $Q^T SQ = D$, where the diagonal matrix $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ and the eigenvalues have been arranged so that $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_m \geq 0$. This choice of Q is precisely the one we were looking for above and we have that $Q^T SQ = D = \frac{1}{N-1}\widehat{Y}\widehat{Y}^T$. We are also in a position to now understand why the eigenvalues of S are nonnegative because for the special choice of $\mathbf{z} = \widehat{\mathbf{q}}_k$ we have that $\lambda_k \geq 0$ as required.

```
% Find the eigenvalues and eigenvectors of S
[Q,Dvec] = eig(S, 'vector')
```

Output

```
Q =

    0.3346    0.8909    0.3071
   -0.9423    0.3189    0.1016
   -0.0074   -0.3234    0.9462
```

```
Dvec =

    6.4680
   23.0978
  280.1806
```

```
% Sort in descending order
[Dvec, perm] = sort(Dvec, 'descend');
Q = Q(:, perm);
D = diag(Dvec);

% Show that Q is orthogonal
Q'*Q

Q*Q'
```

Output

```
ans =

    1.0000    0.0000    0.0000
    0.0000    1.0000    0.0000
    0.0000    0.0000    1.0000

ans =

    1.0000    0.0000    0.0000
    0.0000    1.0000    0.0000
    0.0000    0.0000    1.0000
```

```
% Show that Q^T S Q is a diagonal matrix
Q'*S*Q
```

Output

```
ans =

  280.1806   -0.0000    0.0000
         0   23.0978   -0.0000
    0.0000   -0.0000    6.4680
```

```
% Check the rank of Xhat transpose
rank(Xhat')
```

Output

```
ans =
    3
```

We refer to the eigenvector corresponding to the largest eigenvalue of S the *first principal component* of the data, the eigenvector corresponding to the second largest eigenvalue of S the *second principal component* of the data, and so on. Note that the first principal component represents the new variable y_1 as a linear combination of the original variables $x_1, x_2, \ldots, x_m$ weighted according to the entries in the eigenvector.

PCA becomes extremely valuable when most of the variation in the data is due to only a small number of the new variables $y_1, y_2, \ldots, y_m$. Since $\mathrm{tr}(D) = \mathrm{tr}(Q^T SQ) = \mathrm{tr}(Q^T QS) = \mathrm{tr}(S)$, we see that the *Total Variance* = $\sum_{k=1}^{m} \lambda_k$. Furthermore, since the variance of variable y_k is λ_k, the quotient $\frac{\lambda_k}{\mathrm{tr}(S)}$ measures the fraction of the total variance captured by the new variable y_k.

```
% Calculate the percentage of the variance accounted for by each principal
% component
D(1,1)/trace(S)

D(2,2)/trace(S)

D(3,3)/trace(S)
```

Output

```
ans =

    0.9045
ans =

    0.0746
ans =

    0.0209
```

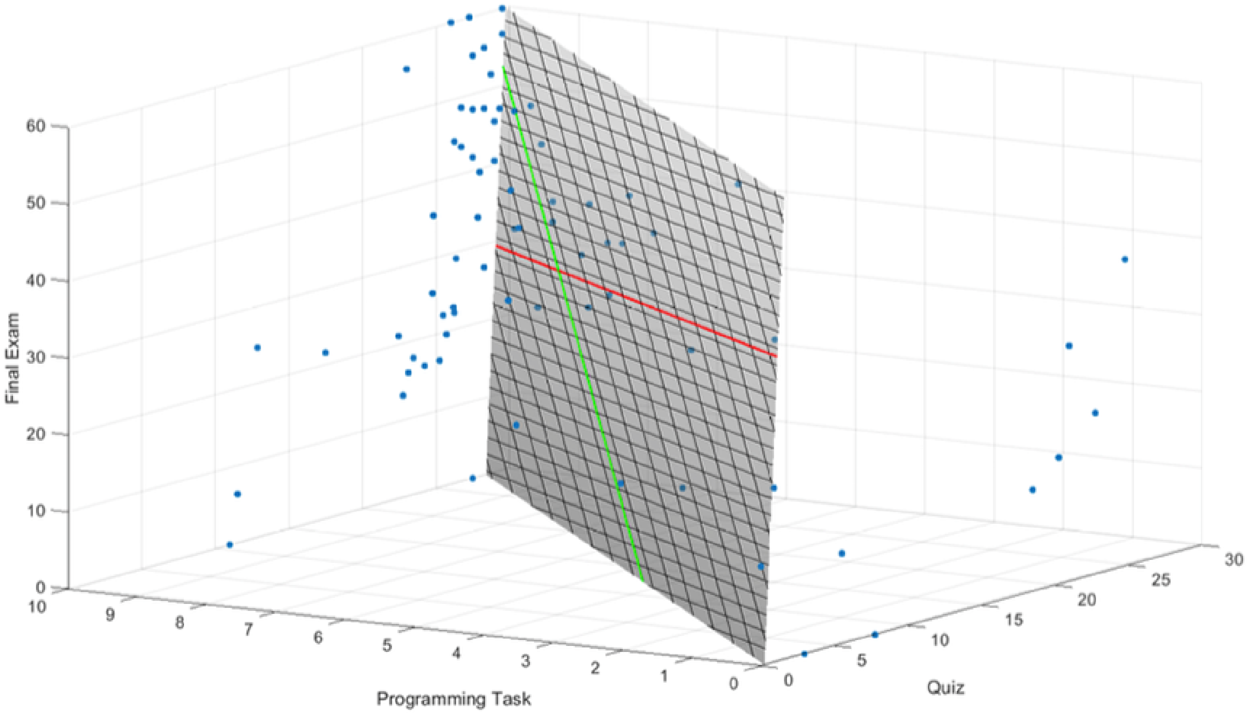
This shows us that almost 98% of the variation in the data is accounted for by the first two principal components. In fact, over 90% of the variation is accounted for by just the first principal component alone. So, depending on what we wanted to do with the data, we could consider storing just the first principal component, instead of all three dimensions of the original data set.

Now let’s plot the original data together with the principal component directions.

```
% Plot the original data with principal component axes and plane
t = [-50 50];
figure;
scatter3(Xn(1,:)',Xn(2,:)',Xn(3,:)', 'filled')
hold on
pc1_line = [sum(Xn,2)./N + t(1)*Q(:, 1), sum(Xn,2)./N + t(end)*Q(:, 1)];
plot3([pc1_line(1, 1) pc1_line(1, 2)], [pc1_line(2, 1) pc1_line(2, 2)],...
    [pc1_line(3, 1) pc1_line(3, 2)], '-g', 'LineWidth', 2);
pc2_line = [sum(Xn,2)./N + t(1)*Q(:, 2), sum(Xn,2)./N + t(end)*Q(:, 2)];
```

```
plot3([pc2_line(1, 1) pc2_line(1, 2)], [pc2_line(2, 1) pc2_line(2, 2)],...
    [pc2_line(3, 1) pc2_line(3, 2)], '-r', 'LineWidth', 2);

% Generate the plane with tangent vectors as first two principal components
Mu = sum(Xn,2)./N;
funx = @(s, t) Mu(1,1) + t*Q(1, 1) + s * Q(1, 2);
funy = @(s, t) Mu(2,1) + t*Q(2, 1) + s * Q(2, 2);
funz = @(s, t) Mu(3,1) + t*Q(3, 1) + s * Q(3, 2);
s = ezsurf(funx, funy, funz, t);
colormap('gray');
alpha(s, '0.5');
axis([0 30 0 10 0 60]);
```



This plane represents the orthogonal regression plane; it minimises the orthogonal distance between the data and the plane.