Multivariate Analysis (Slides 3)

- In today's class we start to introduce Principal Components Analysis (PCA).
- We start with a quick overview, then get into the mechanics of PCA.
- First we'll introduce the sample mean and covariance matrix.
- Then we will re-cap on methods for constrained optimization using Lagrange multipliers.

Motivation

- For data with many variables/dimensions it is often difficult to comprehend or visualize inherent associations.
- There may be multicollinearity: two or more variables are highly correlated.
- PCA can be thought of as a method for re-expressing the data so as to reveal its internal structure and explain its variation through the use of a few linear combinations of the original variables.
- Imagine that a data set is made up of three positively correlated variables.
- Could one linear combination of the three variables be sufficient to convey most of the information given by the three individual variables?
- PCA can be used as either a dimension reduction technique, or as a method for identifying associations among variables.

How PCA works

- The aim of PCA is to describe the variation in a set of *correlated* variables $X_1, X_2, \ldots, X_m$ in terms of a new set of *uncorrelated* variables $Y_1, Y_2, \ldots, Y_p$, where, hopefully, $p << m$, and where each Y_j is a linear combination of the $X_1, X_2, \ldots, X_m$.
- The new 'variables', or *principal components*, are derived in decreasing order of importance so that the first principal component Y_1 accounts for more variation in the original data than any other possible linear combination of $X_1, X_2, \ldots, X_m$.
- The second principal component Y_2 is chosen to account for as much of the remaining variation as possible subject to the constraint that it be uncorrelated with Y_1. And so on...
- The hope is that the first few principal components will account for a substantial amount of the variation in the original data, and as such, can be used as a convenient lower dimensional summary of it.

Sample Mean

- We have a data set consisting of n observations, each consisting of m measurements.
- That is, we have data

$$\begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1m} \end{pmatrix}, \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2m} \end{pmatrix}, \ldots, \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nm} \end{pmatrix}$$

- The sample mean of each variable is $\overline{x}_j = \sum_{i=1}^{n} x_{ij}/n$.

Sample Covariance Matrix

- The sample covariance matrix $\mathbf{Q}$ is the matrix whose terms q_{ij} are defined to be

$$q_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j),$$

  for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, m$.

- The previous results discussed for covariances of random variables also apply for sample covariances.

- We divide by $n - 1$ rather than n because we have estimated the mean of the random variables by the sample mean. If the true $\mathbb{E}[\mathbf{X}]$ were known then we would use the usual formula and divide by n. Effectively we have lost a degree of freedom.

Aside: Why $n-1$ and not n?

- Consider a sample of size n. It can be thought of as existing as a point in an n dimensional space and so has n degrees of freedom in movement. Each sample member is free to take any value it likes. If, however, you fix the sample mean, then the sample values are constrained to have a fixed sum. You can change $n-1$ of the sample members, but the last member must have value to ensure the sample mean is unchanged. We are interested in the deviations from the sample mean, or in fact the product of these deviations, and whilst we may sum over them n times, this is in truth a sum of only $n-1$ independent things.
- Of course we could use an n, but then the distributional assumptions for the sample statistics become much more complicated to express later down the line.

Linear Combinations of Data

- Let Y be a linear combination of the m variables, *i.e.*, $Y = \sum_{j=1}^{m} a_j X_i$.
- Then $Y = \mathbf{a}^T \mathbf{X}$.
- To determine $\mathrm{Var}(Y)$ we would require $\mathrm{Cov}(\mathbf{X})$.
- Instead estimate $\mathrm{Cov}(\mathbf{X})$ through $\mathbf{Q}$.
- Then $\mathrm{Var}(Y)$ is estimated as $\mathbf{a}^T \mathbf{Q} \mathbf{a}$.
- Similarly, if $Z = \mathbf{b}^T \mathbf{X}$, then $\mathrm{Cov}(Y, Z)$ is estimated as $\mathbf{a}^T \mathbf{Q} \mathbf{b} = \mathbf{b}^T \mathbf{Q} \mathbf{a}$.

Data with 7 variables

- The mean of the heptathlon data variables is,

X100m	HighJump	ShotPutt	X200m	LongJump	Javelin	X800m
13.9	1.78	13.8	24.6	6.10	45.0	135.5

- The sample covariance matrix of the heptathlon data is,

	X100m	HighJump	ShotPutt	X200m	LongJump	Javelin	X800m
X100m	0.17	-0.01	-0.02	0.16	-0.09	-0.39	-0.03
HighJump	-0.01	0.01	-0.03	-0.03	0.01	0.05	-0.07
ShotPutt	-0.02	-0.03	1.59	-0.01	0.06	1.42	2.01
X200m	0.16	-0.03	-0.01	0.38	-0.14	-0.60	0.40
LongJump	-0.09	0.01	0.06	-0.14	0.08	0.40	-0.16
Javelin	-0.39	0.05	1.42	-0.60	0.40	15.80	0.47
X800m	-0.03	-0.07	2.01	0.40	-0.16	0.47	24.44

- What is the mean and variance of $8 \times$ (100m time) $+ 4 \times$ (200m time)?

Eigenvalues and Eigenvectors

- The eigenvalues of a sample covariance matrix are non-negative.
- A sample covariance matrix of dimension $m \times m$ has m orthonormal eigenvectors.
- **Example:** The old faithful data has sample covariance matrix

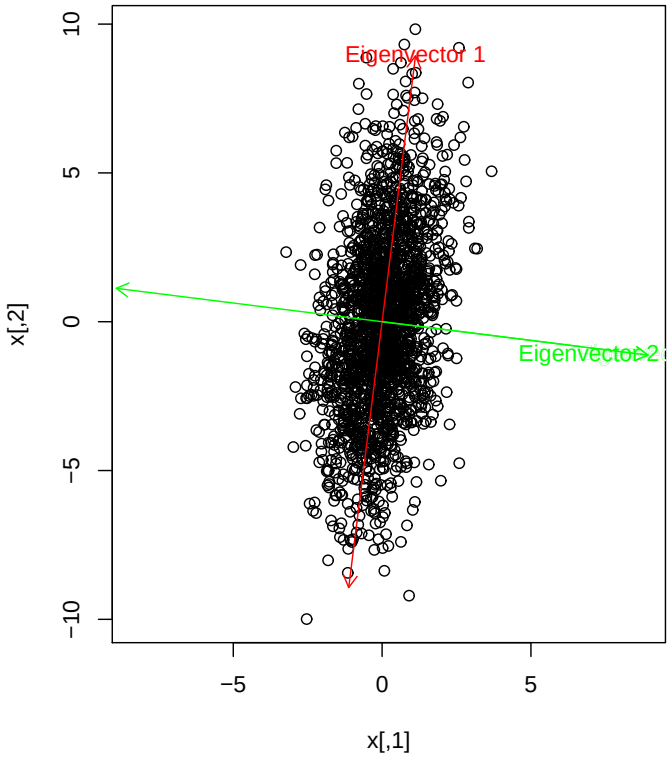
$$\mathbf{Q} = \begin{pmatrix} 1.3 & 14.0 \\ 14.0 & 184.8 \end{pmatrix}.$$

- The eigenvalues of $\mathbf{Q}$ are 185.9 and 0.244.
- The corresponding eigenvectors for these eigenvalues are

$$\begin{pmatrix} 0.0755 \\ 0.9971 \end{pmatrix} \text{ and } \begin{pmatrix} 0.9971 \\ -0.0755 \end{pmatrix}.$$

Sample Covariance Eigenvectors

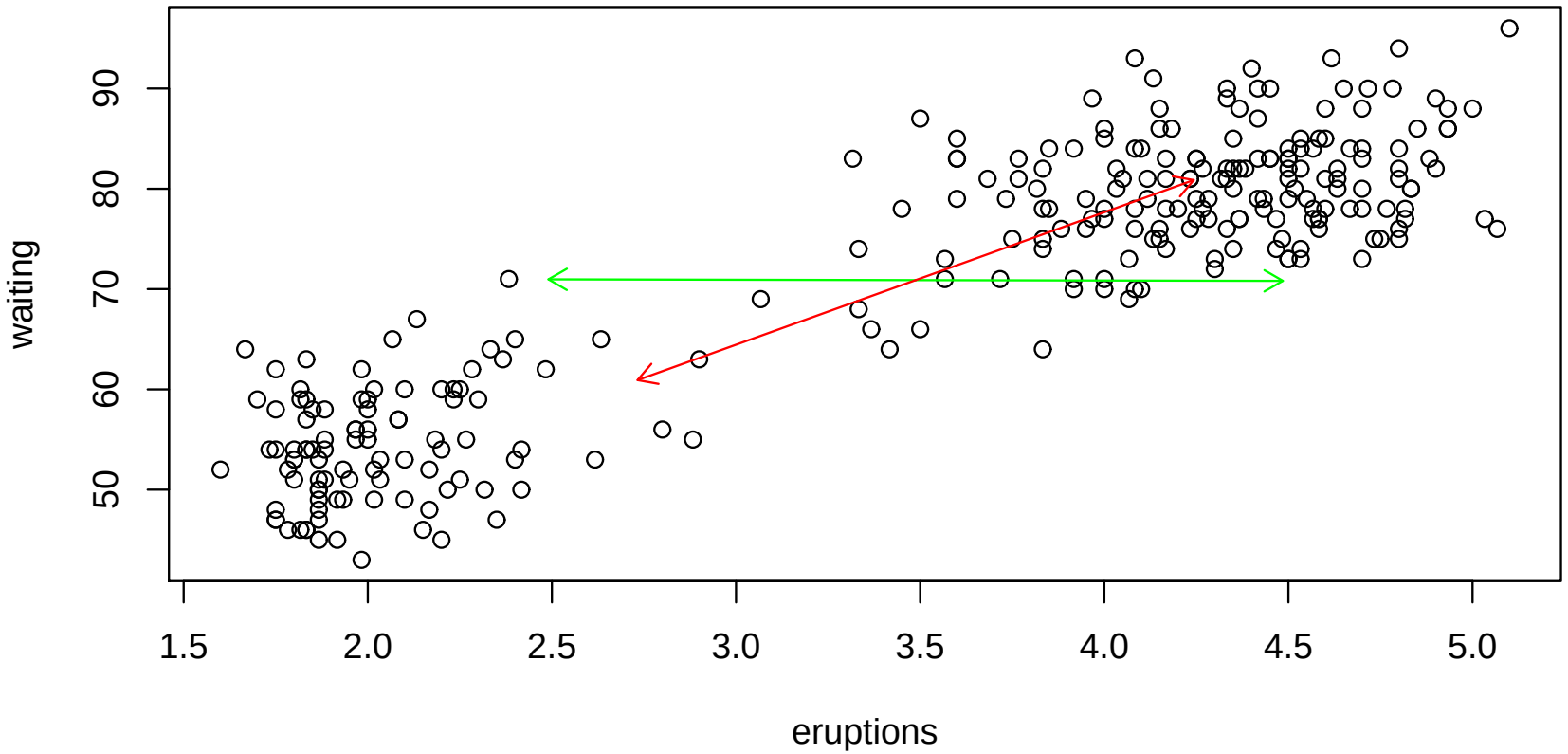
- A sample of 2000 bivariate data values were generated from **Q** with zero mean.
- The **directions** of the two eigenvectors from the zero mean were added.



- One of these follows the long direction of the scatter.

Old Faithful Covariance Eigenvectors

- Arrows in the direction of the eigenvectors were overlaid from the mean of the actual data.
- Beware of the scaling: eigenvectors of a symmetric matrix are orthogonal!



- One of these follows the long direction of scatter.

Heptathlon Data

- More difficult to visualize!
- The eigenvalues of the sample covariance matrix are,

```
24.656 15.936  1.290  0.476  0.082  0.016  0.002
```

- The corresponding unit eigenvectors are,

```
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] -0.002  0.025 -0.016 -0.450  0.832 -0.316 -0.075
[2,] -0.003 -0.004  0.021  0.060  0.007 -0.295  0.953
[3,]  0.091 -0.088 -0.991  0.007  0.000  0.027  0.030
[4,]  0.015  0.040 -0.013 -0.825 -0.525 -0.203 -0.006
[5,] -0.005 -0.026 -0.029  0.332 -0.180 -0.878 -0.291
[6,]  0.066 -0.992  0.094 -0.053  0.005  0.005 -0.001
[7,]  0.994  0.073  0.084  0.016  0.009 -0.006 -0.002
```

- Look down the columns to read off the eigenvectors.

Constrained Optimisation

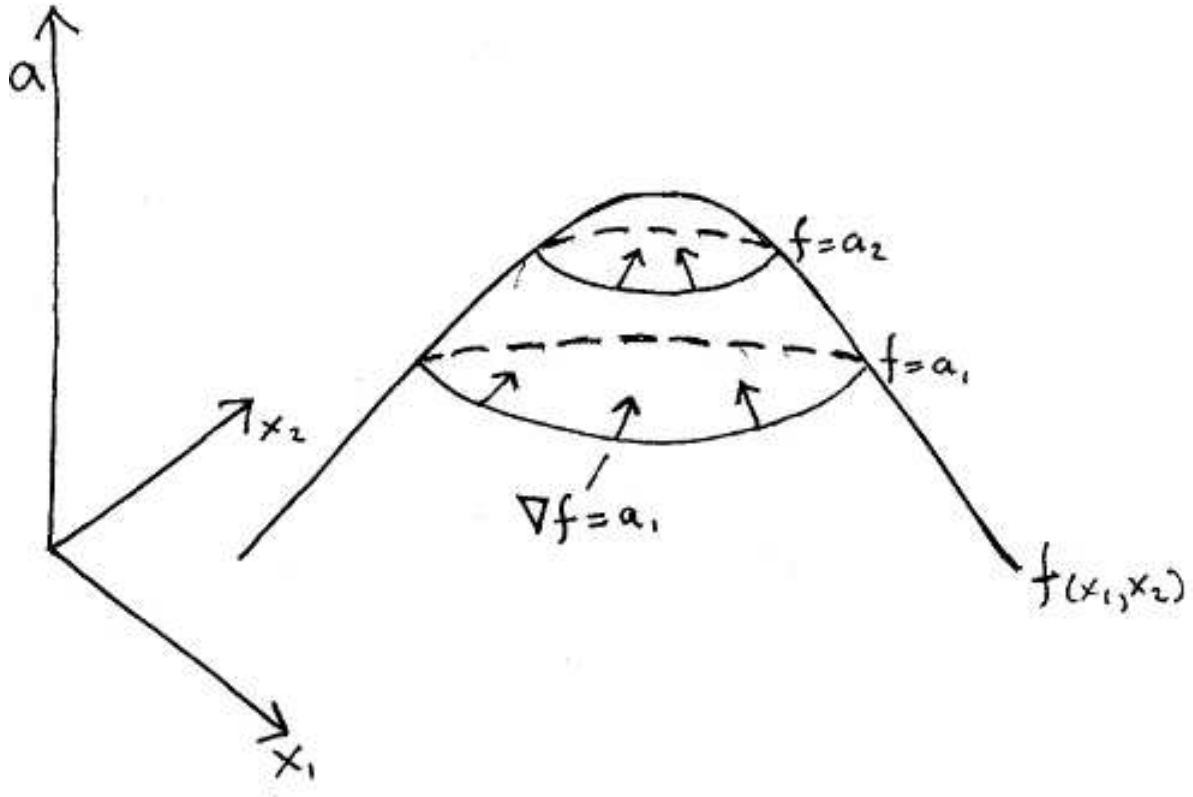
- Consider the following problem.
- **Question:** Which linear combination $\mathbf{a}^T\mathbf{X}$ of the variables in the data has maximum variance, subject to the constraint $\mathbf{a}^T\mathbf{a} = 1$?
- We have seen that the variance of any linear combination $\mathbf{a}^T\mathbf{X}$ of the variables in the data is

$$\text{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{Q}\mathbf{a}.$$

- Hence, we have the following optimisation problem.
- **Problem:** Find $\mathbf{a}$ to maximize $\mathbf{a}^T\mathbf{Q}\mathbf{a}$ subject to $\mathbf{a}^T\mathbf{a} = 1$.
- **Question:** How do we solve this optimization problem?

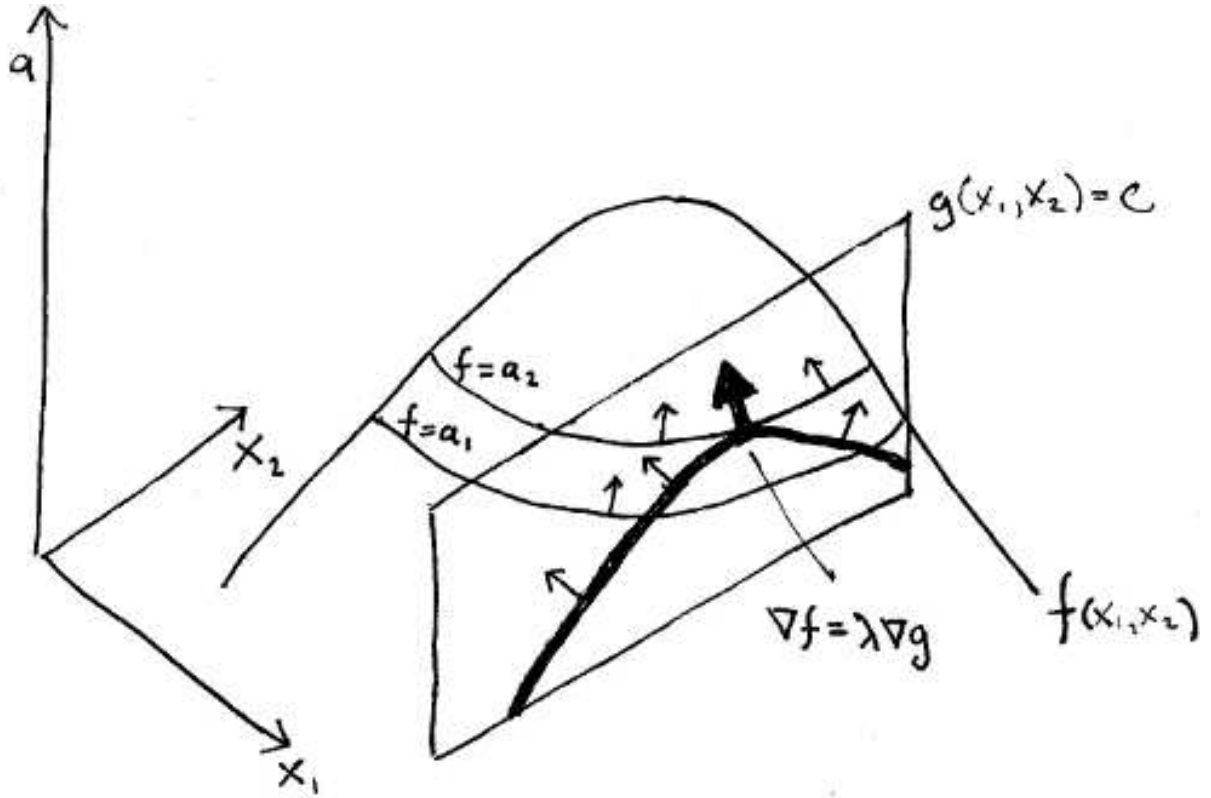
Lagrange Multipliers

- **Swapping notation**
- Given a function $f(\mathbf{x})$, the gradient of f, ∇f (the collection of partial derivatives) indicates the direction of steepest slope. Level curves (lines where $f(\mathbf{x})$ is constant) run perpendicular to the gradient.



Lagrange Multipliers

- A constraint $g(\mathbf{x}) = c$ represents a plane that cuts through the variable space. When the direction of the gradient of the constraint equals the direction of the gradient of the f function, then that location constitutes a local maxima (the actual gradients point in the same direction and differ at most by a scalar coefficient).



Lagrange Multipliers

- Hence we want the location where $\nabla f = \lambda \nabla g$.
- In other words, find $\mathbf{x}$ so that $\nabla f - \lambda \nabla g = \mathbf{0}$.
- This gives us $m + 1$ equations with $m + 1$ unknowns.

Lagrange Multipliers

- We can use the method of Lagrange multipliers to find the maximum value of $\mathbf{a}^T\mathbf{Q}\mathbf{a}$ subject to the constraint that $\mathbf{a}^T\mathbf{a} - 1 = 0$.
- First, let $P = \mathbf{a}^T\mathbf{Q}\mathbf{a} - \lambda(\mathbf{a}^T\mathbf{a} - 1)$.
- Our solution is found by computing partial derivatives $\frac{\partial P}{\partial a_k}$, for $k = 1, 2, \ldots, m$, and $\frac{\partial P}{\partial \lambda}$.
- We then seek to solve these $m + 1$ equations when they are set to equal zero.
- Notice that $\frac{\partial P}{\partial \lambda} = -(\mathbf{a}^T\mathbf{a} - 1)$, so solving this equation guarantees that the constraint is satisfied.

Lagrange Multipliers

- **Hint:** Many people find it easier to switch out of matrix notation when doing the differentiation here (but if you can do it in vector notation the solution is much quicker).
- We can write $P = \sum_{i=1}^{m} a_i^2 q_{ii} + \sum_{i=1}^{m} \sum_{j \neq i}^{m} a_i a_j q_{ij} - \lambda \left(\sum_{i=1}^{m} a_i^2 - 1 \right)$. Hence,

$$\begin{aligned} \frac{\partial P}{\partial a_k} &= 2a_k q_{kk} + \sum_{j \neq k} a_j q_{kj} + \sum_{i \neq k} a_i q_{ik} - 2\lambda a_k \\ &= 2a_k q_{kk} + 2 \sum_{i \neq k} a_i q_{ik} - 2\lambda a_k \end{aligned}$$

- So,

$$\frac{\partial P}{\partial a_k} = 0 \Rightarrow \sum_{i=1}^{m} a_i q_{ik} - \lambda a_k = 0$$

Lagrange Multipliers

- We have seen that, for $k = 1, \ldots, m$, $\frac{\partial P}{\partial a_k} = 0$ implies:

$$(a_1, a_2, \ldots, a_m) \begin{pmatrix} q_{1k} \\ q_{2k} \\ \vdots \\ q_{mk} \end{pmatrix} = \lambda a_k$$

- This can be expressed more compactly using matrices:

$$(a_1, a_2, \ldots, a_m) \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \cdots & q_{mm} \end{pmatrix} = \lambda (a_1, a_2, \ldots, a_m)$$

Lagrange Multipliers

- In matrix notation, we have

$$\mathbf{a}^T\mathbf{Q} = \lambda\mathbf{a}^T$$

- Taking the transpose of both sides gives,

$$\mathbf{Q}^T\mathbf{a} = \mathbf{Q}\mathbf{a} = \lambda\mathbf{a}$$

- That is, $\mathbf{a}$ is an eigenvector of $\mathbf{Q}$. The constraint also tells us that it's a unit eigenvector.
- Finally, due to the maximization part of the problem, we also know it is the eigenvector with largest eigenvalue.

$$\text{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{Q}\mathbf{a} = \mathbf{a}^T\lambda\mathbf{a} = \lambda\mathbf{a}^T\mathbf{a} = \lambda$$

Principal Components

- The first principal component of the data set is the linear combination of the variables that has greatest variance.
- This corresponds to taking a linear combination of the variables where the weights are as given by the eigenvector of $\mathbf{Q}$ with largest eigenvalue. This eigenvalue also represents the variance of the linear combination.
- What about other principal components?
- The second principal component is the linear combination of the variables where the weights are as given by the eigenvector of $\mathbf{Q}$ corresponding to the second largest eigenvalue. Again the eigenvalue represents the variance of this linear combination. And so on...
- All principal components are orthogonal to one another.

Summary

- A principal components analysis constructs a set of linear combinations of the data variables.
- Subsequent principal components have smaller and smaller variances.
- All principal components are uncorrelated with each other.