Principal Components Analysis (PCA)

Visualization

Dimension Reduction

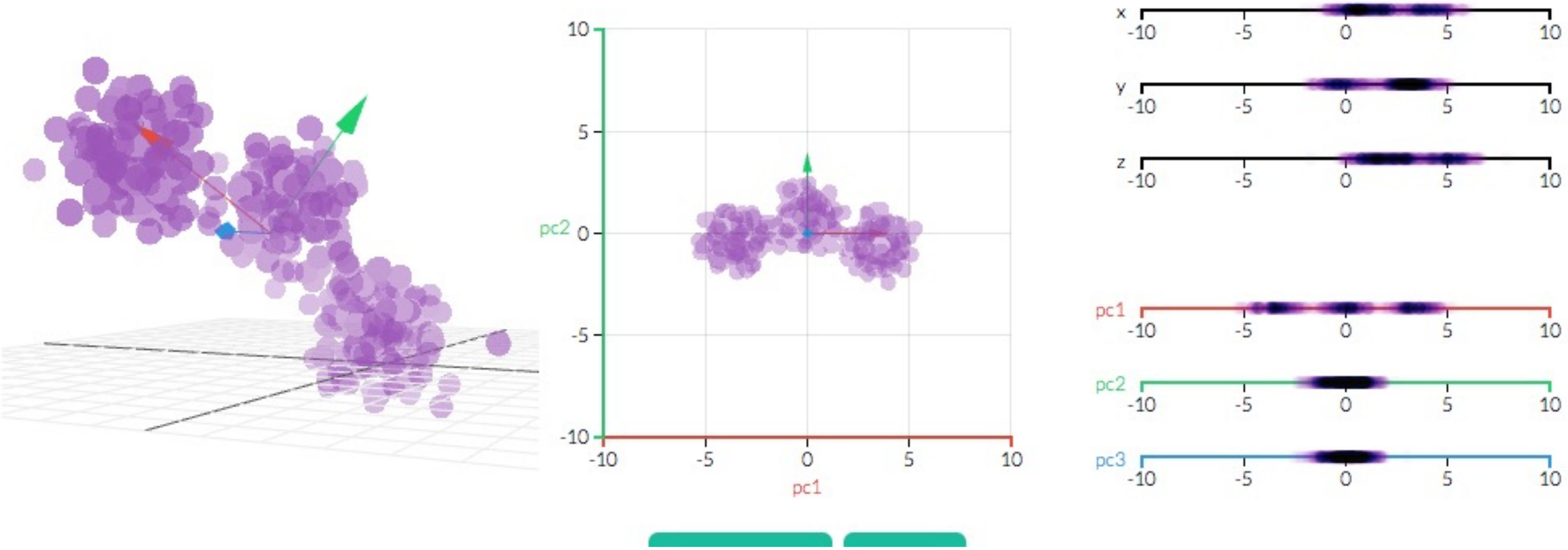
References

- Theory: M. Richardson, ”Principal Component Analysis,” from http://www.dsc.ufcg.edu.br/~hmg/disciplinas/posgraduacao/rn-copin-2014.3/material/SignalProcPCA.pdf
- Visualizations: PCA Victor Powell and Lewis Lehe, http://setosa.io/ev/principal-component-analysis/

PCA

- Technique used to emphasize variation and bring out strong patterns in a dataset.
- Often used to make data easy to explore and visualize.
- Transforms a number of possibly correlated variables into a smaller number of variables called principal components.

Example

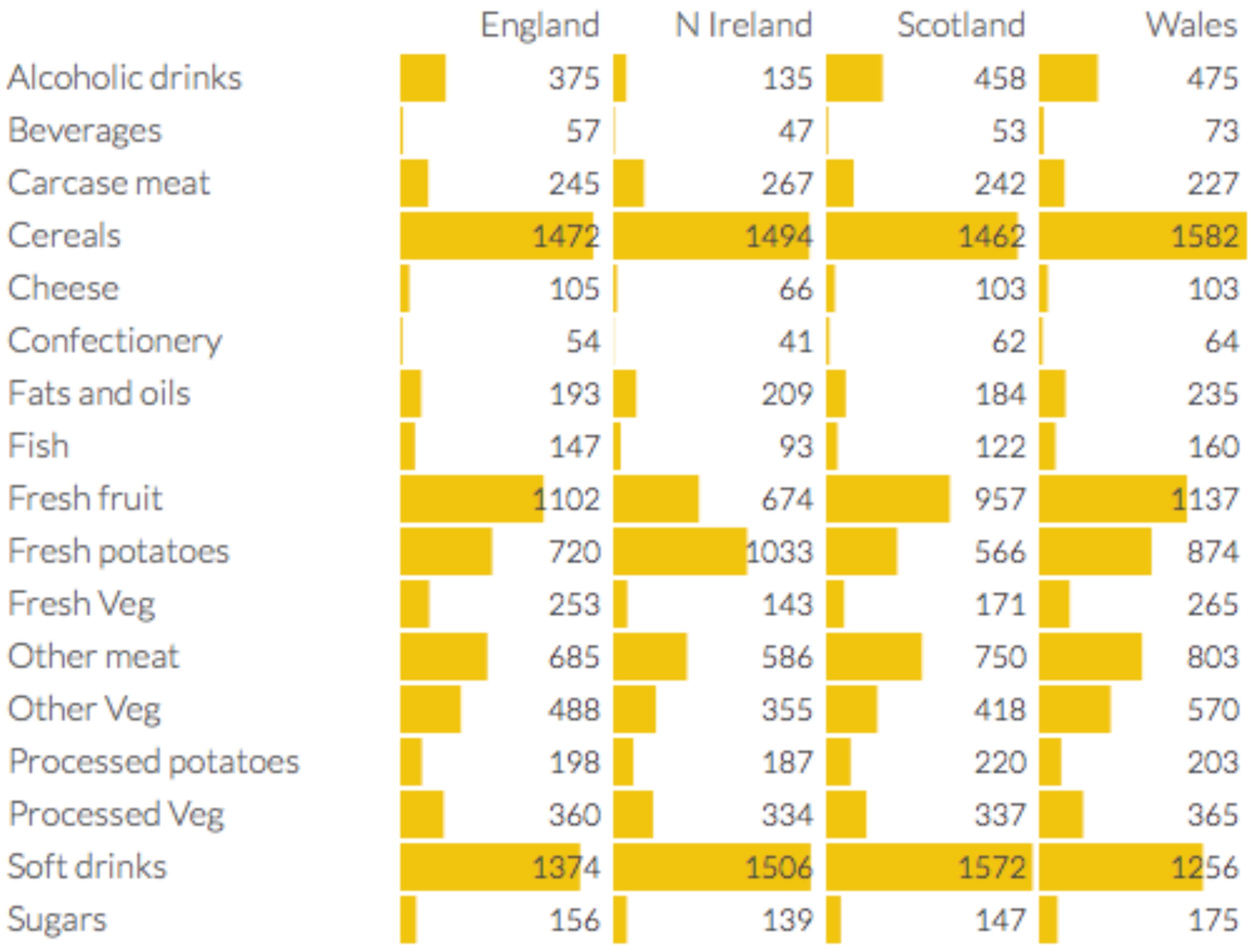


http://setosa.io/ev/principal-component-analysis/

Eating in the UK

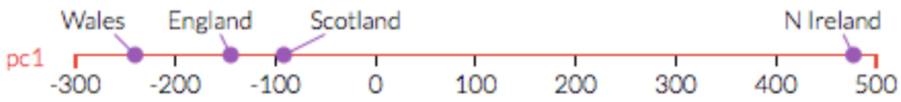
17 Dimensions

How do countries differ?



	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Principal Components (PCs)



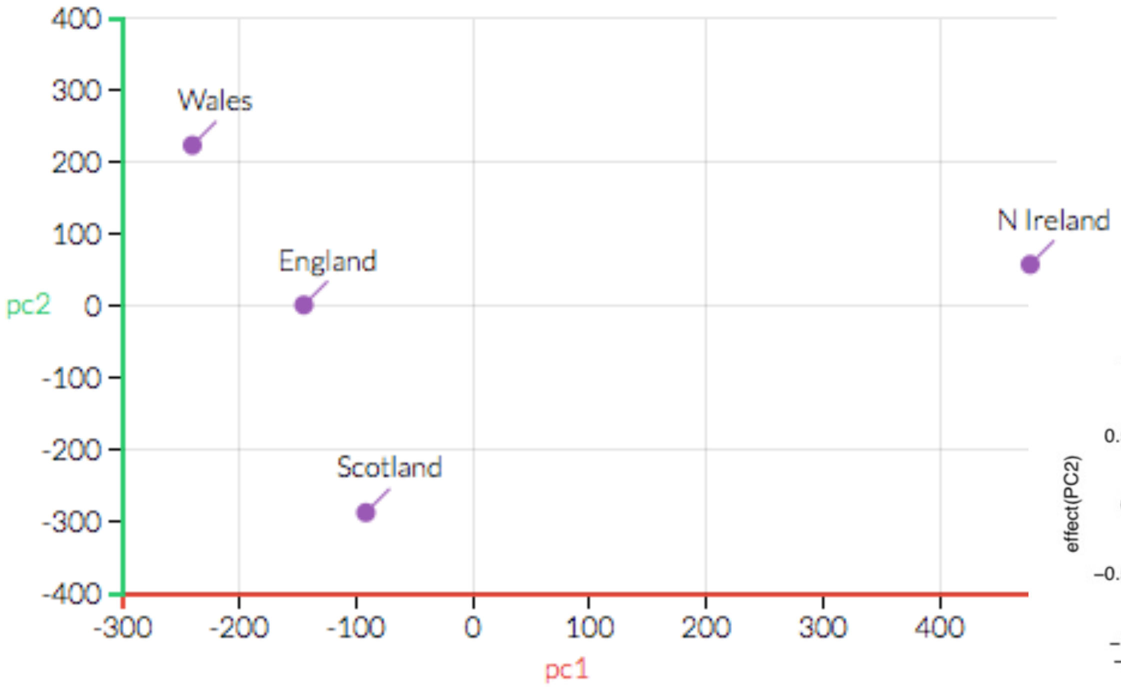


Figure 3: Eigenspectrum

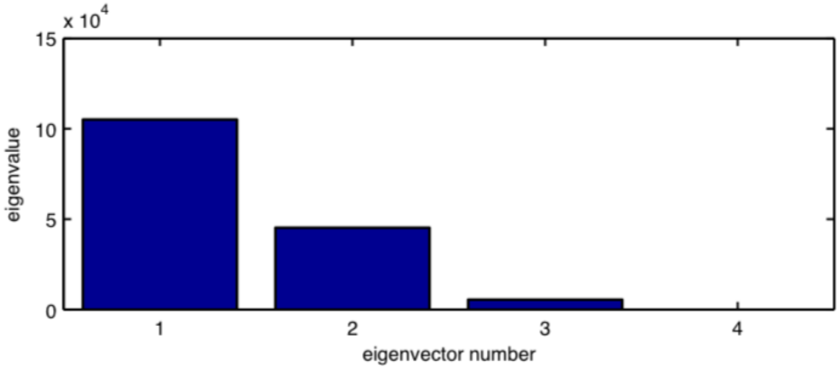
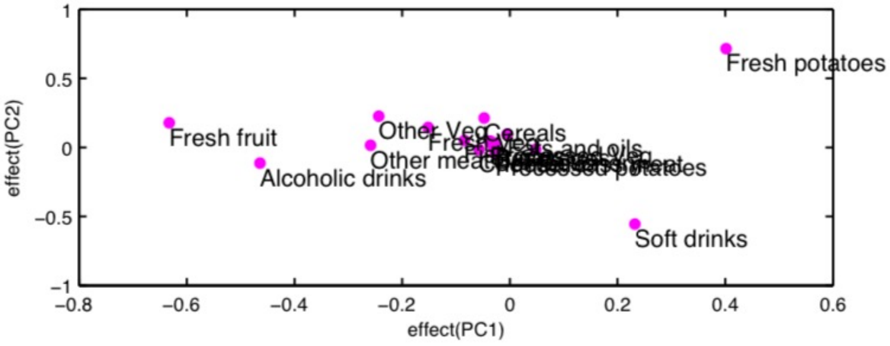


Figure 4: Load plot

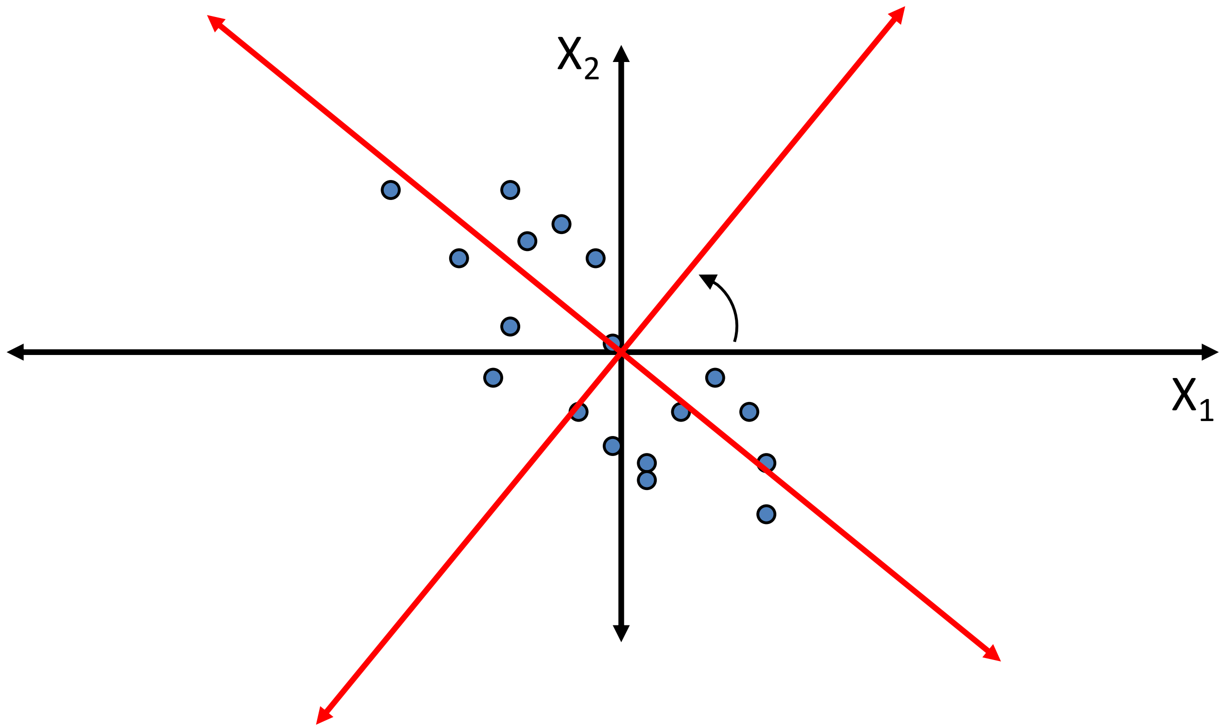


Goal: Project data to lower dimensional space

- PCA finds a new coordinate system in which every point has a new (x,y) value.
- Axes don't actually mean anything physical; they’re linear combinations of the variables called "principal components" that are chosen to give one axes lots of variation.

Trick: Rotate Coordinate Axes

Suppose we have a population measured on m random variables $X_1,\ldots,X_m$. Note that these random variables represent the p-axes of the Cartesian coordinate system. Our goal is to develop a new set of m axes (linear combinations of the original m axes) in the directions of greatest variability:



This is accomplished by rotating the axes.

Data Matrix

- Rows are features that we measure
- Columns are measurements of these features
- Size N features for rows, n measurements

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iN} \end{bmatrix} \qquad \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{i=n} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iN} \end{bmatrix}$$

Covariance

- Variance and Covariance are a measure of the “spread” of a set of points around their center of mass (mean)
- **Variance** – measure of the deviation from the mean for points in one dimension e.g., potato eating
- **Covariance** as a measure of how much each of the dimensions vary from the mean with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & grades obtained.
- The covariance between one dimension and itself is the variance

Covariance

$$\text{covariance}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

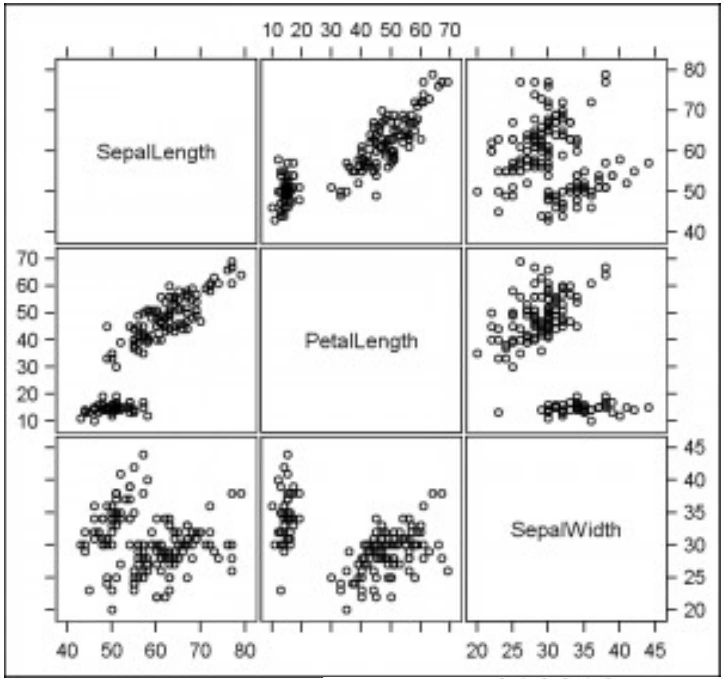
- Given a 3-dimensional data set (x,y,z)
- Measure the covariance between the x and y dimensions, the y and z dimensions, and the x and z dimensions.
- Measuring the covariance between x and x , or y and y , or z and z gives the variance of the x , y and z dimensions respectively.

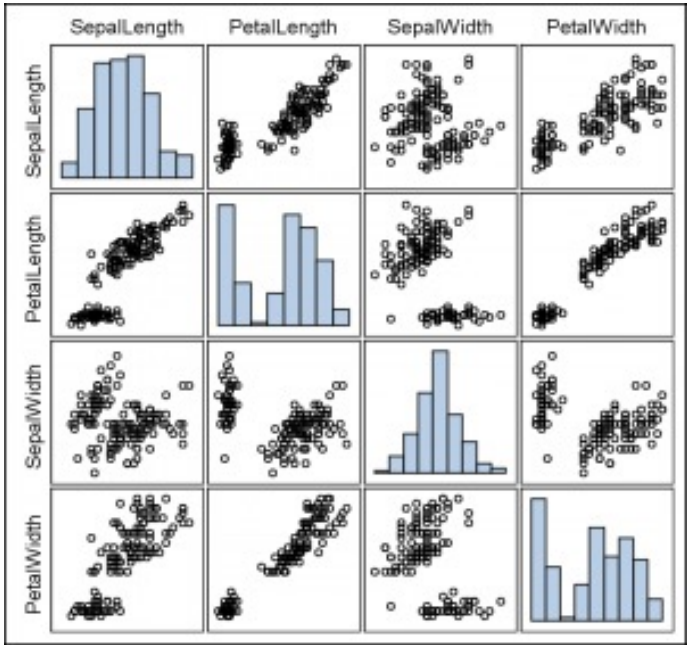
$$C = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{bmatrix}$$

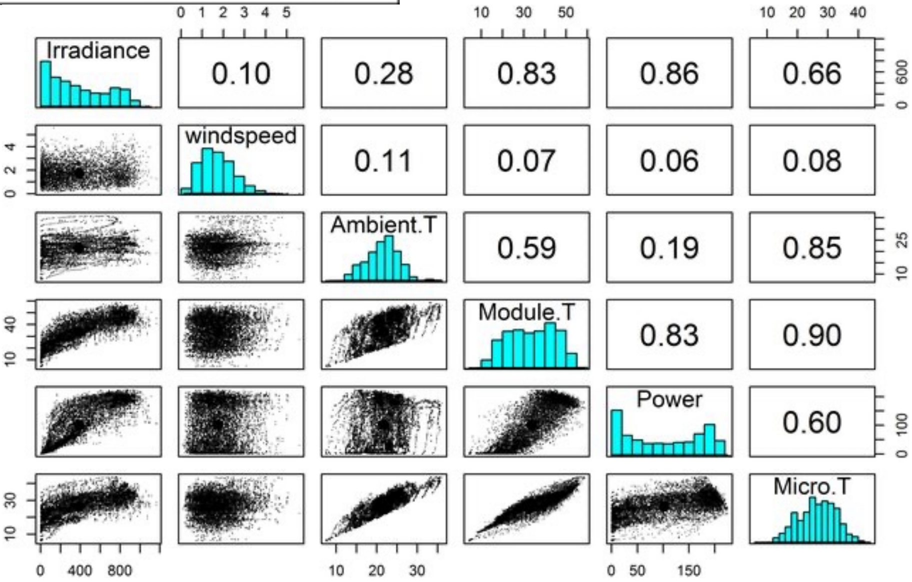
Covariance Matrix

- Diagonal is the variances of x_1, x_2
- cov(x,y) = cov(y,x) hence matrix is symmetrical about the diagonal
- N-dimensional feature data will result in NxN covariance matrix
- Exact value is not as important as the sign.
- Positive value of covariance indicates both dimensions increase or decrease together
- Negative value indicates while one increases the other decreases, or vice-versa
- Zero covariance is zero: the two dimensions are independent of each other

Scatterplot Matrix







Matrix Computation of Covariance

- Form X as the N x n matrix with columns,
- $x_1 - \bar{x}, x_2 - \bar{x}, \ldots$
- $X = [x_1 - \bar{x} |\ x_2 - \bar{x} |\ \ldots x_n - \bar{x},]$
- Note: subtracting the mean is equivalent to translating the coordinate system to the location of the mean.

Matrix Computation of Covariance

- Let Q = X X^T be the N x N matrix:

$$Q = XX^T = \begin{bmatrix} \mathbf{x}_1 - \bar{\mathbf{x}} & \mathbf{x}_2 - \bar{\mathbf{x}} & \cdots & \mathbf{x}_n - \bar{\mathbf{x}} \end{bmatrix} \begin{bmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^T \end{bmatrix}$$

Notes:

1. Q is square
2. Q is symmetric
3. Q is the covariance matrix [aka scatter matrix]
4. Q can be very large (in bioinformatics, N is often the number of genes or gene products!)

Mathematical Preliminaries

- Eigenvectors
  - A eigenvector of a linear transformation is a non-zero vector that does not change its direction

$$\mathbf{Ae} = \lambda \mathbf{e}$$

  - Found by solving the Characteristic Polynomial (n-th order)

$$\det\left(\mathbf{A} - \lambda \mathbf{I}\right) = 0$$

Symmetric matrices

- All eigenvalues are real.
- Eigenvectors are orthogonal

PCA

- Orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (first principal component), second greatest variance on the second coordinate, and so on.
- Problem Formulation:
  - Data Matrix **X** (Nxn), column wise sample means are zero, columns are repetitions, rows are features
  - Find a family of weight vectors that maximizes the variance

PCA

- Define independence by considering the variance of the data in the original basis.
- Seeks to de-correlate the original data by finding the directions in which variance is maximized.

$$\sigma_Z^2 = E[(Z - \mu)^2]$$

$$\sigma_{\mathbf{r}}^2 = \frac{1}{n}\mathbf{r}\mathbf{r}^T$$

Derivation

- Define the i-th column of X: $\mathbf{x}_{(i)}$ is one data point with n measurements

- Find weights (Loadings) $\mathbf{w}_{(k)} = \left(w_1, \ldots, w_p\right)_{(k)}$
- Weights map each row vector of X into a new vector of scores, $\mathbf{t}_{(i)} = \left(t_1, \ldots, t_k\right)_{(i)}$

$$t_{k(i)} = \mathbf{x}_{(i)}\mathbf{w}_{(k)}$$

PCA Derivation

- The loading vector has to maximize the variance:

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \mathbf{w} \right)^2 \right\}$$

- Can expand into matrix form:

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\mathbf{X}\mathbf{w}\|^2 \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right\}$$

- For a positive semidefinite matrix such as $\mathbf{X}^T\mathbf{X}$**,** the quotient's maximum possible value is the largest **eigenvalue** of the matrix, which occurs when ***w*** is the corresponding eigenvector.

PCA Derivation kth Component

- Kth components can be found by subtracting the first k-1 principal components from X

$$\mathbf{X}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^T$$

- Then finding the new loading vector, w, which extracts the maximum variance

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \left\| \mathbf{X}_k \mathbf{w} \right\|^2 \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}_k^T \mathbf{X}_k \mathbf{w} \right\}$$

Notes

- $\mathbf{X}^T\mathbf{X}$ is proportional to sample covariance matrix of data set **X**
- Sample covariance between two different principal components is proportional to eigenvalue of $\mathbf{X}^T\mathbf{X}$

$$\begin{aligned} Q(\mathrm{PC}_{(j)}, \mathrm{PC}_{(k)}) &\propto (\mathbf{X}\mathbf{w}_{(j)})^T \cdot (\mathbf{X}\mathbf{w}_{(k)}) \\ &= \mathbf{w}_{(j)}^T \mathbf{X}^T \mathbf{X} \mathbf{w}_{(k)} \\ &= \mathbf{w}_{(j)}^T \lambda_{(k)} \mathbf{w}_{(k)} \\ &= \lambda_{(k)} \mathbf{w}_{(j)}^T \mathbf{w}_{(k)} \end{aligned}$$

A 2D Numerical Example

PCA Example –STEP 1

- Subtract the mean from each of the data dimensions. All the x values have mean(x) subtracted and y values have mean(y) subtracted from them. This produces a data set whose mean is zero.

- Subtracting the mean makes variance and covariance calculation easier by simplifying their equations. The variance and co-variance values are not affected by the mean value.

PCA Example –STEP 1

http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf

DATA:

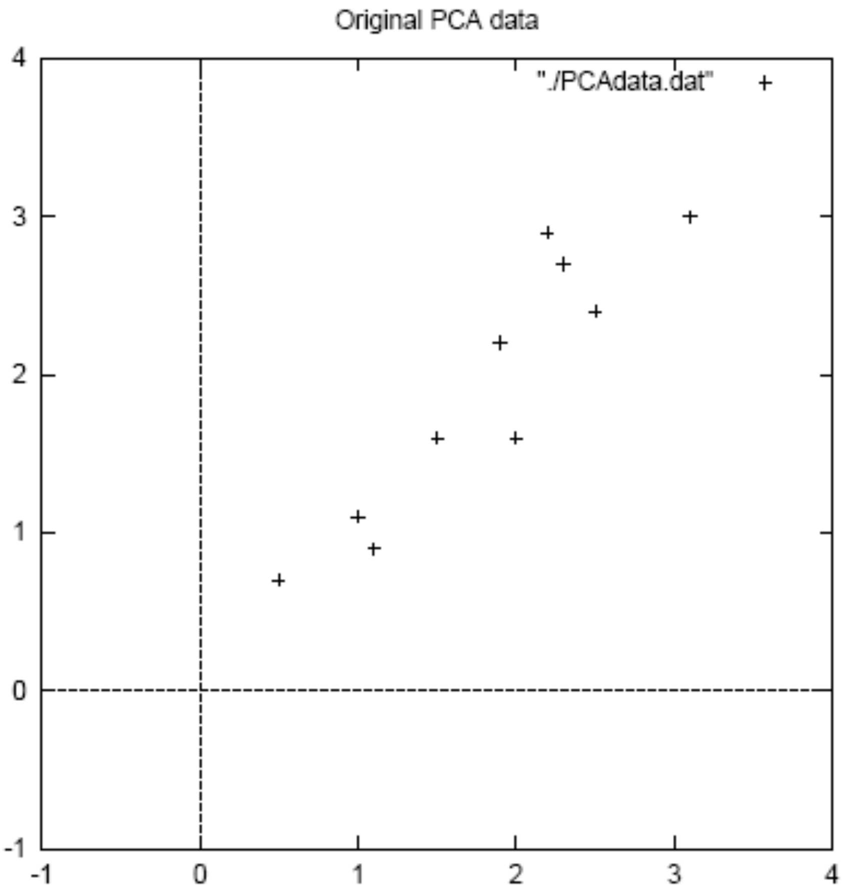
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

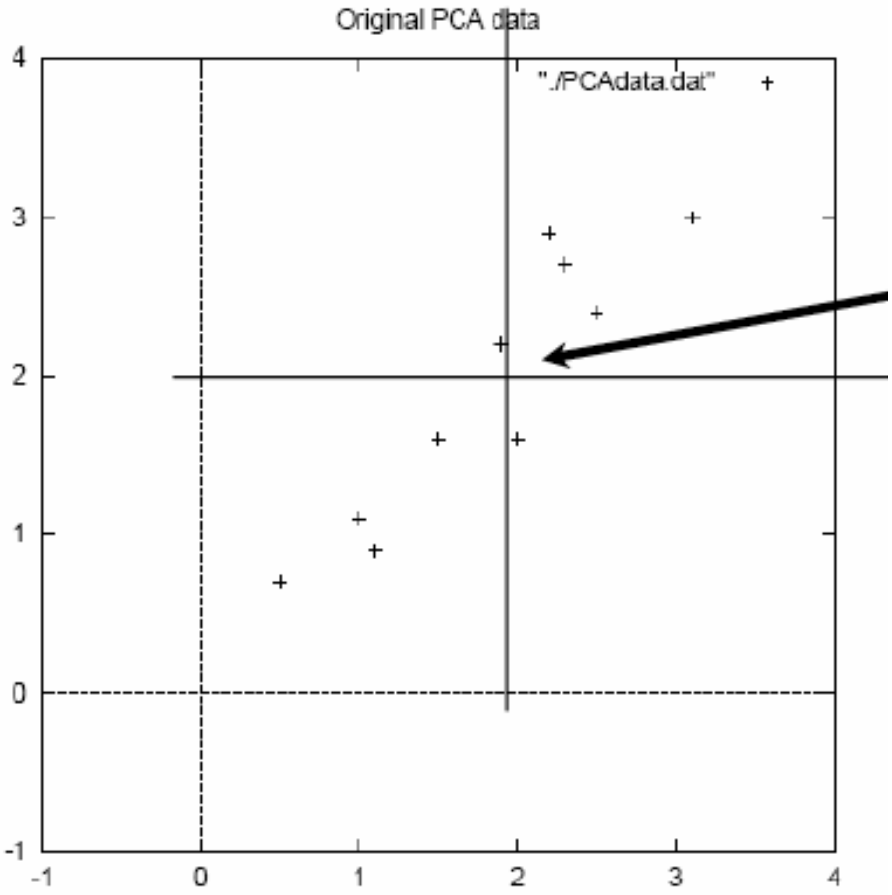
ZERO MEAN DATA:

x	y
.69	.49
-1.3	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

PCA Example –STEP 1

http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf





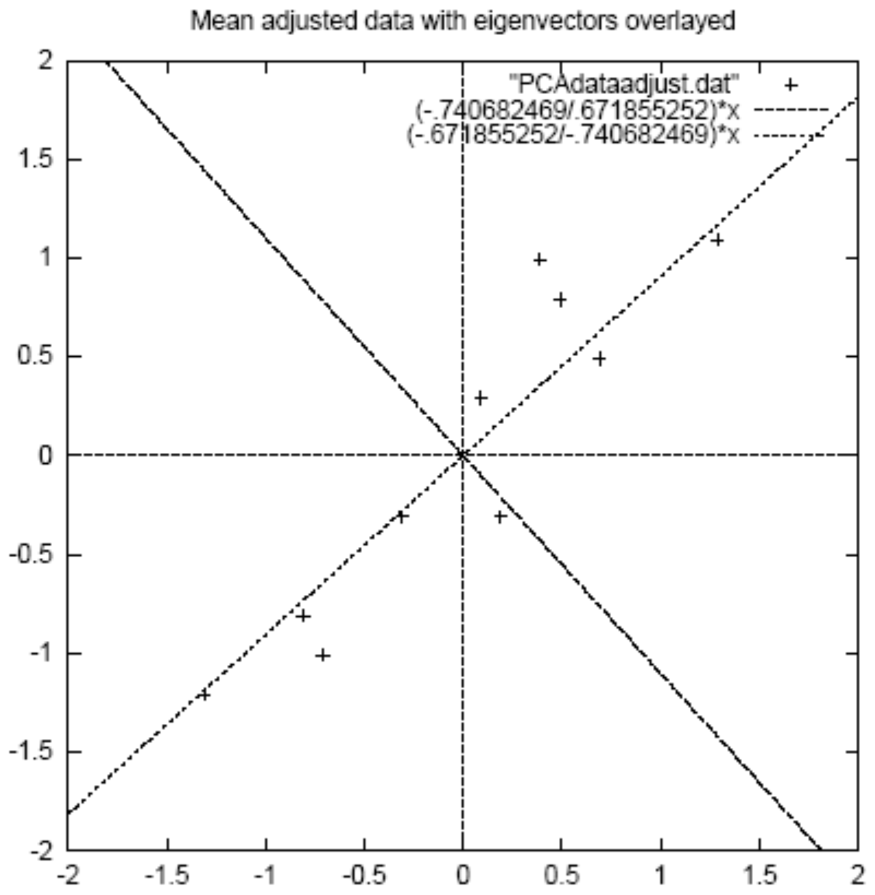
Mean

PCA Example –STEP 2

- Calculate the covariance matrix

  cov = $\begin{pmatrix} .616 & .615 \\ .615 & .717 \end{pmatrix}$

- Since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.



PCA Example –STEP 3

- Calculate the eigenvectors and eigenvalues of the covariance matrix

$$\text{eigenvalues} = \begin{pmatrix} 0.049 \\ 1.284 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.735 & -.678 \\ .678 & -.735 \end{pmatrix}$$

PCA Example –STEP 3

http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf

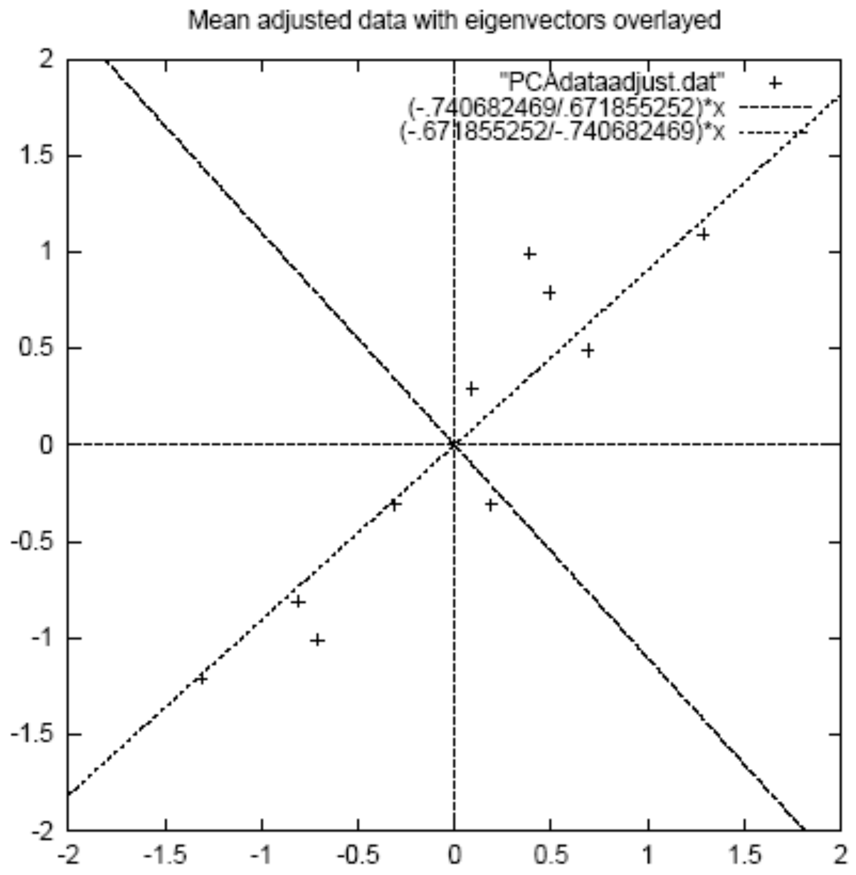


Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

- eigenvectors are plotted as diagonal dotted lines on the plot.
- Note they are perpendicular to each other.
- Note one of the eigenvectors goes through the middle of the points, like drawing a line of best fit.
- The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

PCA Example –STEP 4

- Reduce dimensionality and form *feature vector*

the eigenvector with the *highest* eigenvalue is the **first** *principal component* of the data set.

In our example, the eigenvector with the largest eigenvalue was the one that pointed down the middle of the data.

Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance.

PCA Example –STEP 4

Now, if you like, you can decide to *ignore* the components of lesser significance.

You do lose some information, but if the eigenvalues are small, you don' t lose much

- N dimensions in your data
- calculate N eigenvectors and eigenvalues
- choose only the first p eigenvectors
- final data set has only p dimensions.

PCA Example –STEP 4

- Feature Vector

FeatureVector = $(eig_1\ eig_2\ eig_3 \ldots eig_n)$

We can either form a feature vector using both of the eigenvectors:

$$\begin{pmatrix} -.678 & -.735 \\ -.735 & .678 \end{pmatrix}$$

or, we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

PCA Example –STEP 5

- Deriving the new data

FinalData = RowFeatureVector x RowZeroMeanData

RowFeatureVector is the matrix with the eigenvectors in the columns *transposed* so that the eigenvectors are now in the rows, with the most significant eigenvector at the top

RowZeroMeanData is the mean-adjusted data *transposed*, ie. the data items are in each column, with each row holding a separate dimension.

PCA Example –STEP 5

FinalData transpose: dimensions along columns

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

PCA Example –STEP 5

http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf

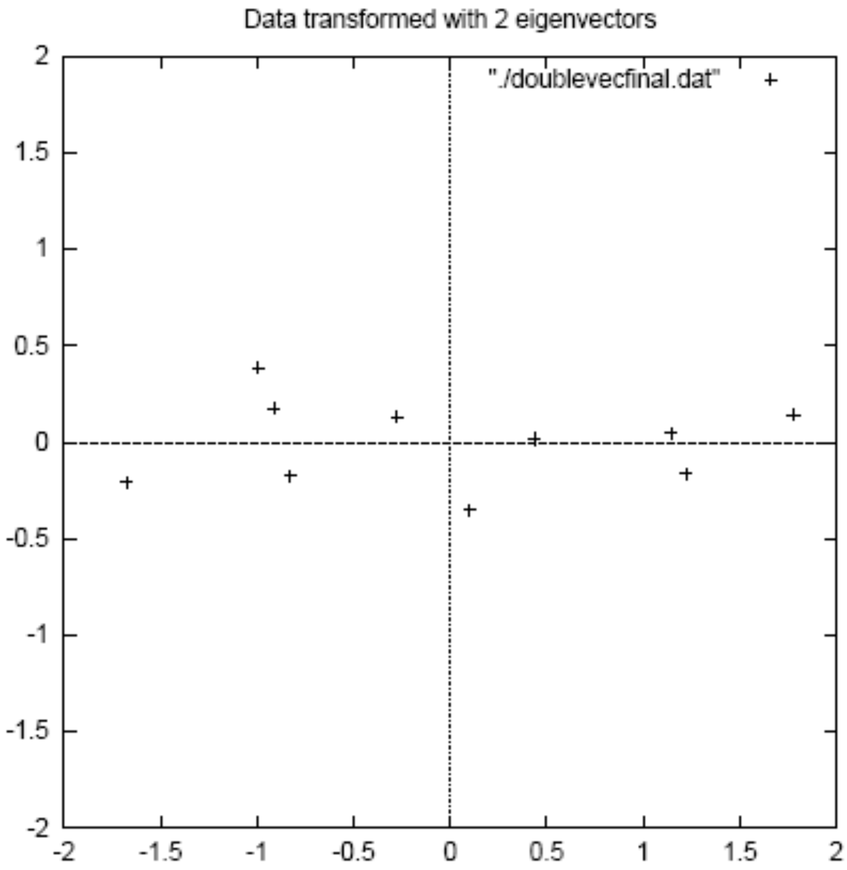


Figure 3.3: The table of data by applying the PCA analysis using both eigenvectors, and a plot of the new data points.

Reconstruction of original Data

- If we reduced the dimensionality, obviously, when reconstructing the data we would lose those dimensions we chose to discard. In our example let us assume that we considered only the x dimension…

Reconstruction of original Data

http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

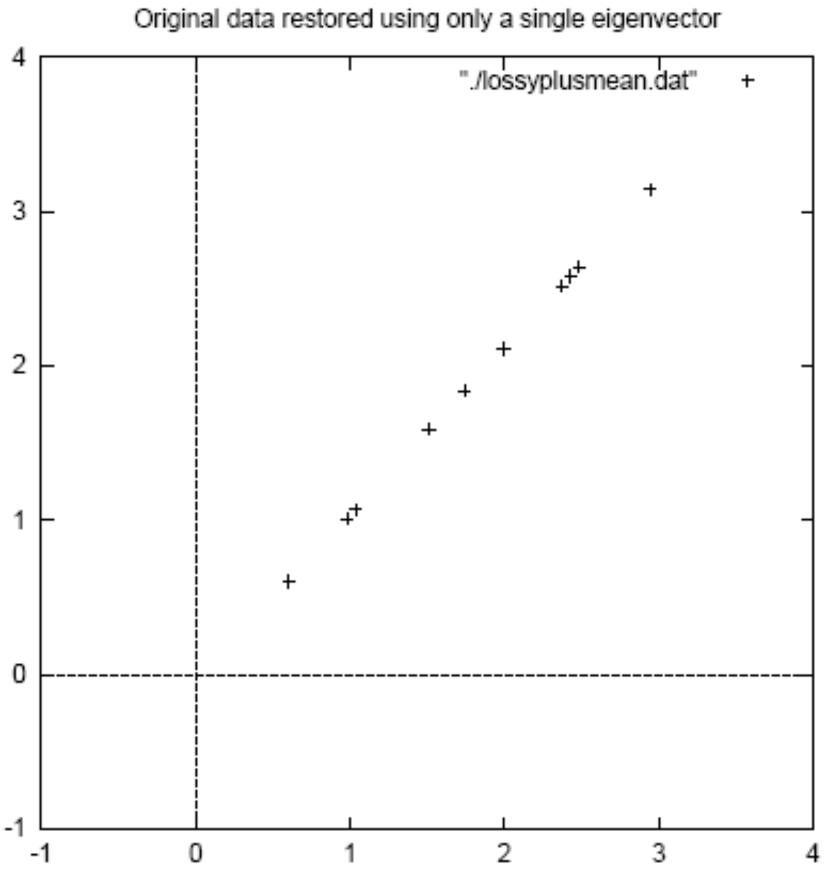


Figure 3.5: The reconstruction from the data that was derived using only a single eigenvector

PCA is sensitive to scale

- If you multiply one variable by a scalar you get different results

(can you show it?)

- This is because it uses covariance matrix (and not correlation)
- PCA should be applied on data that have approximately the same scale in each variable

Normalized (standardized) PCA

- If variables have very heterogenous variances, then standardize them
- The standardized variables X_i*

$$X_i{*}= (X_i\text{-}mean)/\,\overline{SUMvariance}$$

- The new variables all have the same variance (1), so each variable has similar weight.

Interpretation of PCA

- The new variables (PCs) have a variance equal to their corresponding eigenvalue

  $Var(Y_i)=\sum_i \lambda_i$ for all $i=1\ldots p$

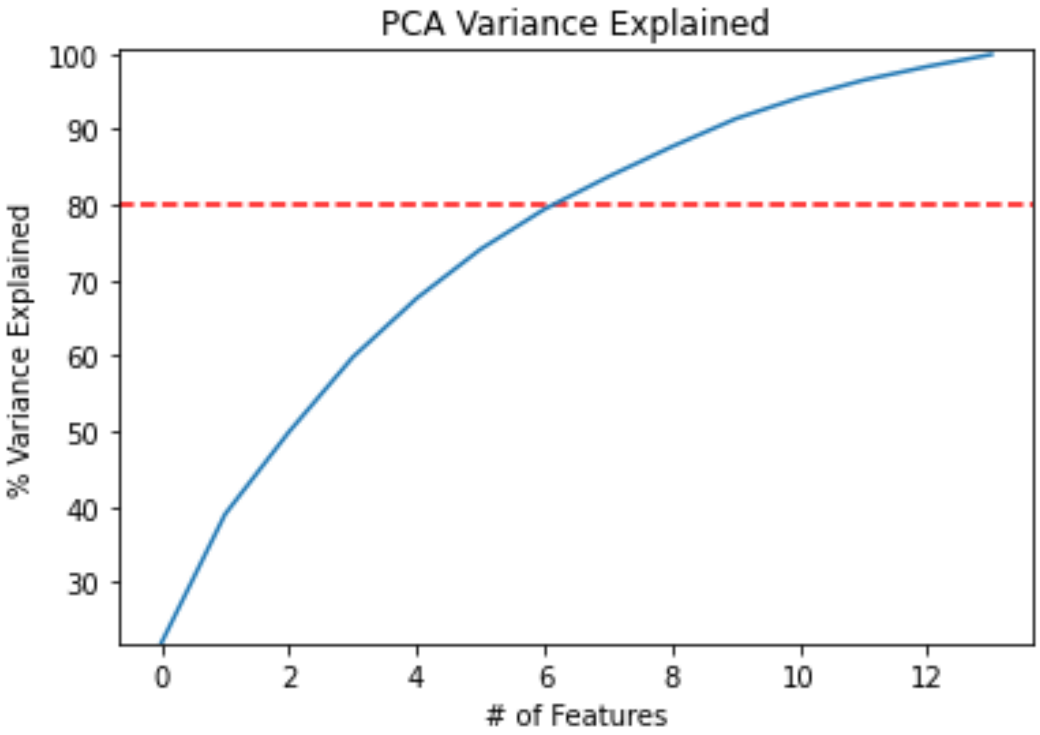
- Small $\sum_i \lambda_i$ ⇔ small variance ⇔ data change little in the direction of component Y_i
- The relative variance explained by each PC is given by $\lambda_i \Big/ \sum_i \lambda_i$

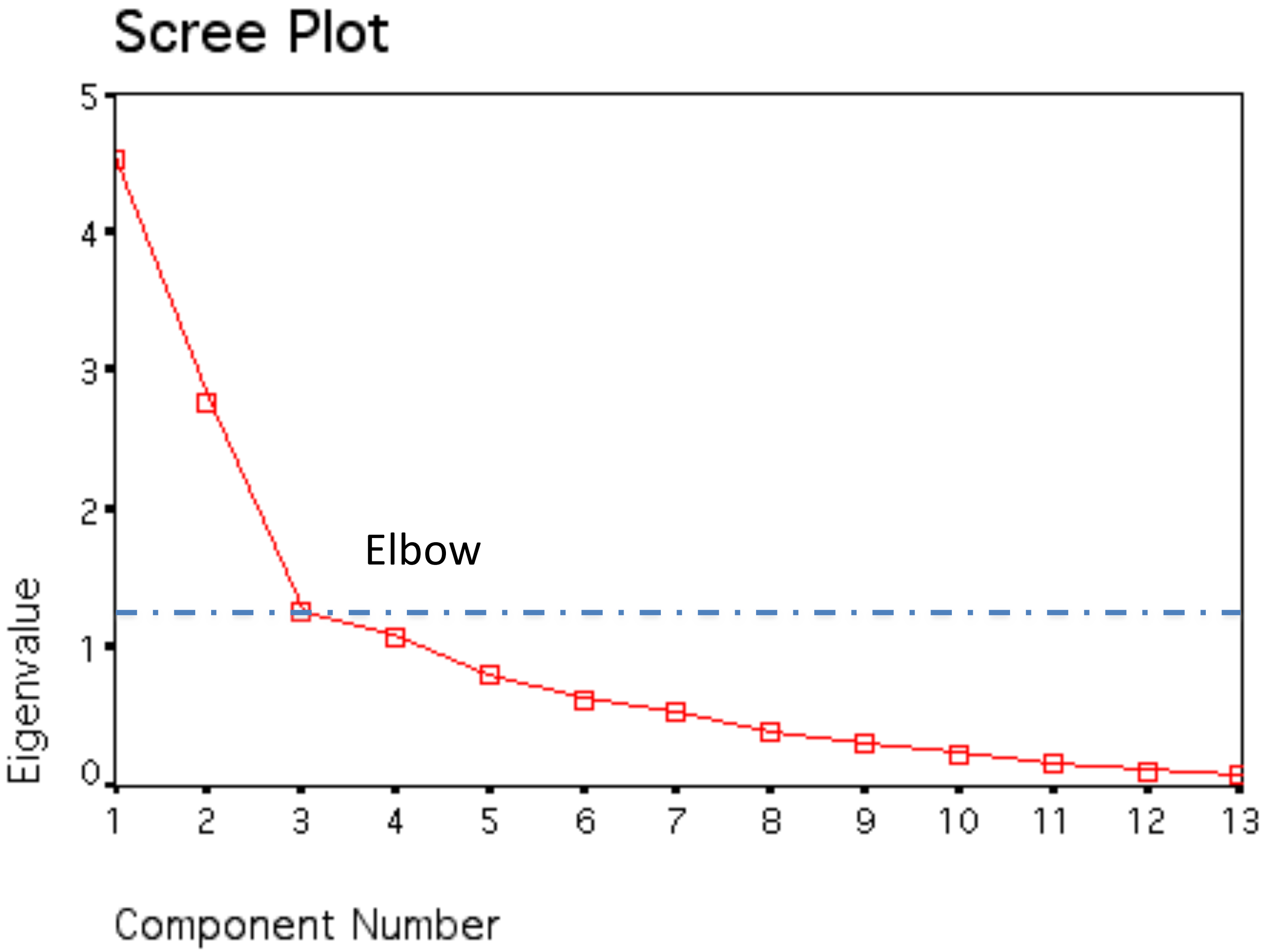
How many components to keep?

- Enough PCs to have a cumulative variance explained by the PCs that is >50-70%, problem dependent
- Kaiser criterion: keep PCs with eigenvalues >1
- Scree plot: represents the ability of PCs to explain variation in data

Percent of variation explained

Rank eigenvalues in order
Sum all eigenvalues = sum of the variance
% of variation explained by 1 PC λ_1/Total
% explained by first two components ($\lambda1+\lambda2$)/Total




Scree Plot
5
4
3
2
1
0
Eigenvalue
1
2
3
4
5
6
7
8
9
10
11
12
13
Component Number

Elbow

Loadings Plot

- A loading plot shows how strongly each characteristic influences a principal component.
  - When two vectors are close, forming a small angle, the two variables they represent are positively correlated.
  - If they meet each other at 90°, they are not likely to be correlated.
  - When they diverge and form a large angle (close to 180°), they are negative correlated.

Figure 4: Load plot

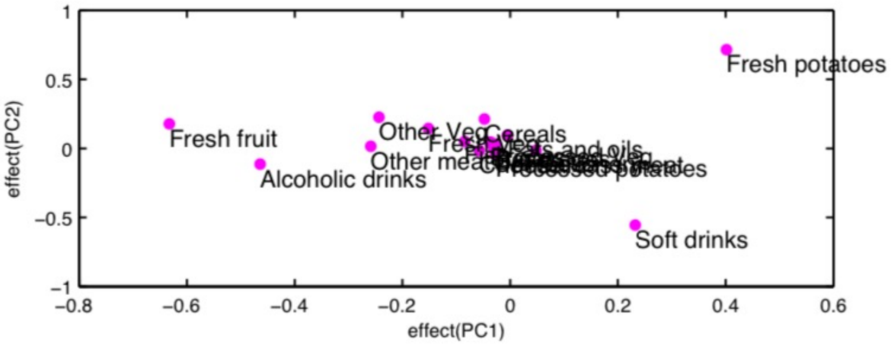
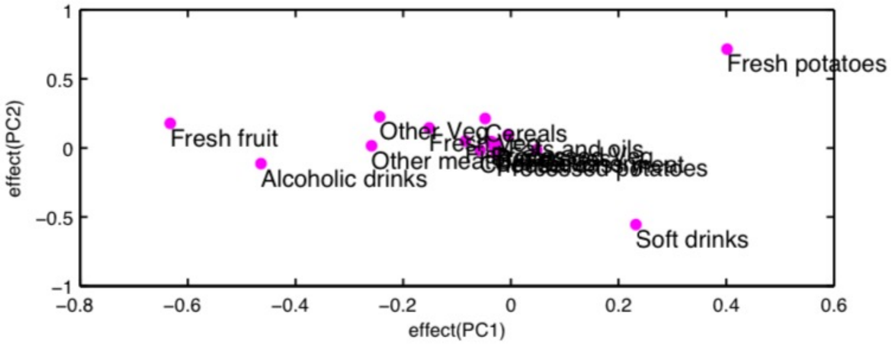
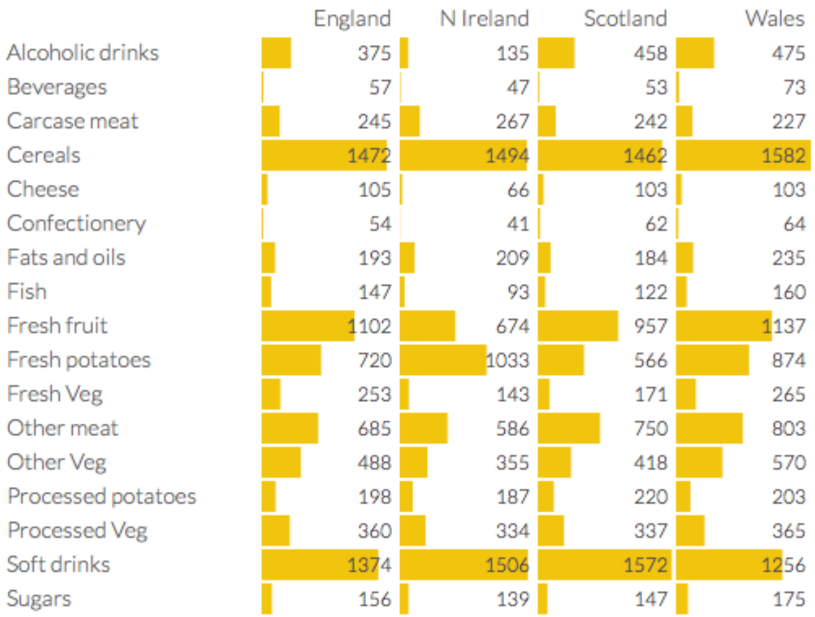


Figure 4: Load plot

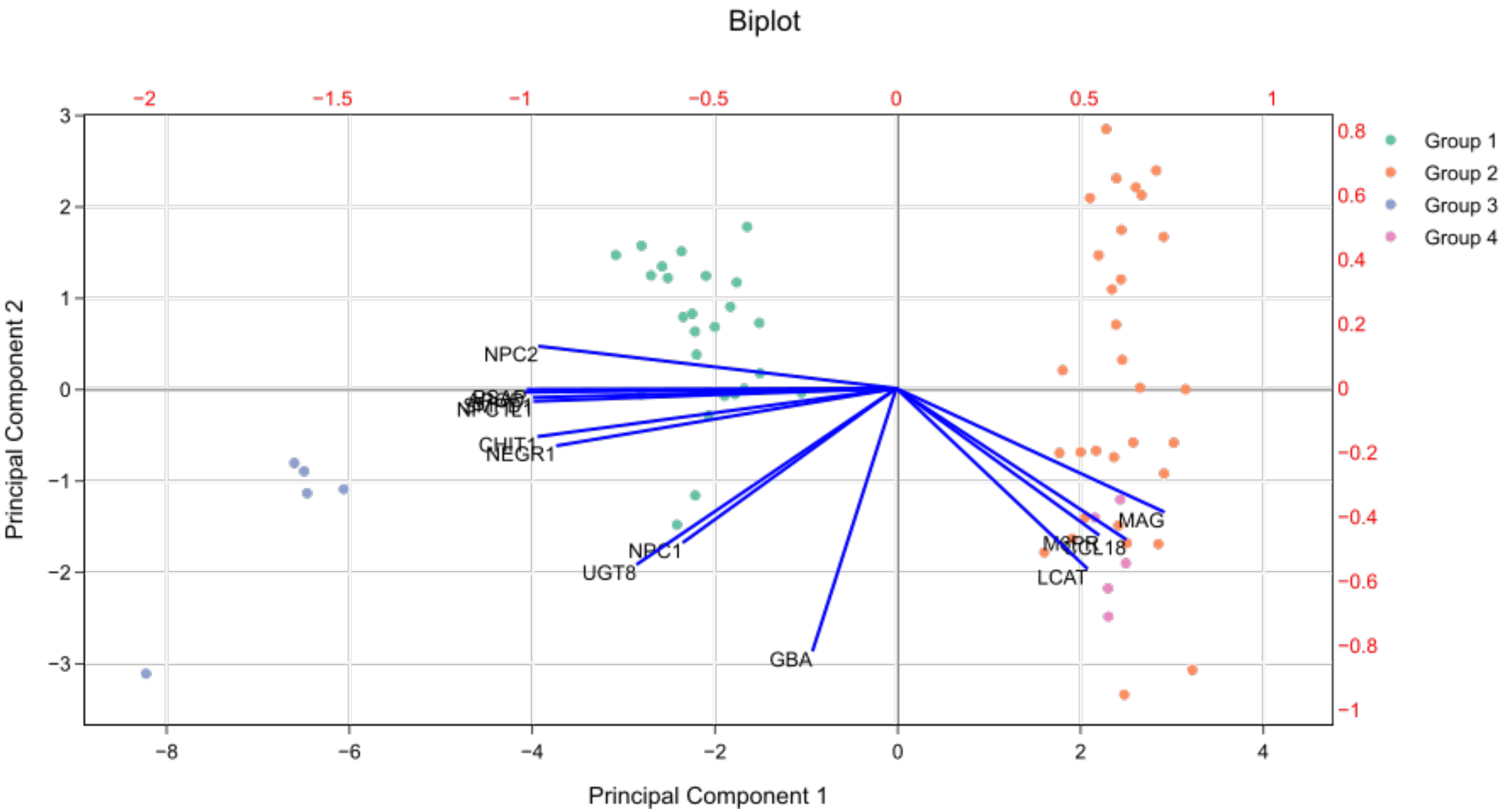




	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

PCA biplot = PCA score plot + loading plot

- PCA biplot merges a PCA plot with a plot of loadings.
- The arrangement is:
  - Bottom axis: PC1 score.
  - Left axis: PC2 score.
  - Top axis: loadings on PC1.
  - Right axis: loadings on PC2.
- The left and bottom axes are of the PCA plot — use them to read PCA scores of the samples (dots).
- The top and right axes belong to the loading plot — use them to read how strongly each characteristic (vector) influence the principal components.


Biplot
Principal Component 1
Principal Component 2
−8
−6
−4
−2
0
2
4
3
2
1
0
−1
−2
−3
−2
−1.5
−1
−0.5
0
0.5
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
Group 1
Group 2
Group 3
Group 4
NPC2
PSAP
NPC1L1
CHIT1
NEGR1
NPC1
UGT8
GBA
MAG
CCL18
LCAT

Interpretation of components

- Look at weights of variables in each component
- If $Y_1= 0.89\ X_1 +0.15X_2-0.77X_3+0.51X_4$
- Then X_1 *and* X_3 have the highest weights and are the most important variable in the first PC
- See the correlation between variables X_i and PCs: circle of correlation

Problems

- Maps to a linear combination of original features, difficult to interpret.
- Often used to show separation between samples/conditions
- Dimension reduction