Advanced Topics in Software Systems
Multimedia data mining and retrieval
The University of Arizona C Sc 630
Spring 2003

Principal Component Analysis

Initialization

```
Off[SetDelayed::write]
Off[General::spell1]
Needs["Statistics`ContinuousDistributions`"]
```

Context: the problem before PCA

Where it starts

- **data samples from a population**
  - **heights,weights and ages for the class**
  - **percentage changes of agricultrural values for counties in Ireland**
  - **acceleration,top speed,and power for a set of cars**

Definitions

- **random variable**
  - **single random variable (100 meter times xor new house square feet xor ...)**
    - **mean, standard deviation**
  - **vector (multi−dimensional)random variable**
    - **means, autocovariance, higher−orderstatistics**

- **variable, axis, dimension**
  - **a single component of a vector random variable**

- **projection**
  - **value of a variable in a defined space**

- **mean**
  - **a measure of the average for a variable**

■ **standard deviation, variance**

a measure of the dispersion for a variable

■ **covariance**

a measure of the degree to which two variables vary together

Example data set

■ **Import some data: name, sex, age, height, weight**

```
(dataset = Import["htwt", "Table"]) // TableForm
```

Aubrey	M	41	74	170
Ron	M	42	68	166
Carl	M	32	70	155
Antonio	M	39	72	167
Deborah	F	30	66	124
Jacqueline	F	33	66	115
Helen	F	26	64	121
David	M	30	71	158
James	M	53	72	175
Michael	M	32	69	143
Ruth	F	47	69	139
Joel	M	34	72	163
Donna	F	23	62	98
Roger	M	36	75	160
Yao	M	.	70	145
Elizabeth	F	31	67	135
Tim	M	29	71	176
Susan	F	28	65	131

■ **Extract height and weight from the data set**

We have a two dimensional random variable

The two axes of variability are height and weight

```
(htwt = dataset[[All, {4, 5}]]) // TableForm
```

```
74      170
68      166
70      155
72      167
66      124
66      115
64      121
71      158
72      175
69      143
69      139
72      163
62      98
75      160
70      145
67      135
71      176
65      131
```

Typical analysis

■ **mean**

```
({mht, mwt} = Sum[htwt[[i]], {i, Length[htwt]}] / Length[htwt]) // N
```

```
{69.0556, 146.722}
```

■ autocovariance matrix

E[[x –E[x]] [x = E[x]]T] ie. the expected value of the products of the variables

variances along the diagonal

covariances of variables i, j in i, j

Normalize so mean is zero for both variables

```
(nhtwt = Table[{htwt[[i, 1]] - mht, htwt[[i, 2]] - mwt},
    {i, Length[htwt]}]) // TableForm
```

```
4.94444         23.2778
-1.05556        19.2778
0.944444        8.27778
2.94444         20.2778
-3.05556        -22.7222
-3.05556        -31.7222
-5.05556        -25.7222
1.94444         11.2778
2.94444         28.2778
-0.0555556      -3.72222
-0.0555556      -7.72222
2.94444         16.2778
-7.05556        -48.7222
5.94444         13.2778
0.944444        -1.72222
-2.05556        -11.7222
1.94444         29.2778
-4.05556        -15.7222
```

autocovariance matrix

```
(acvhtwt = Sum[Dot[Transpose[{nhtwt[[i]]}], {nhtwt[[i]]}],
     {i, Length[htwt]}] / Length[htwt]) // TableForm
```

```
11.7191      65.4599
65.4599      479.867
```

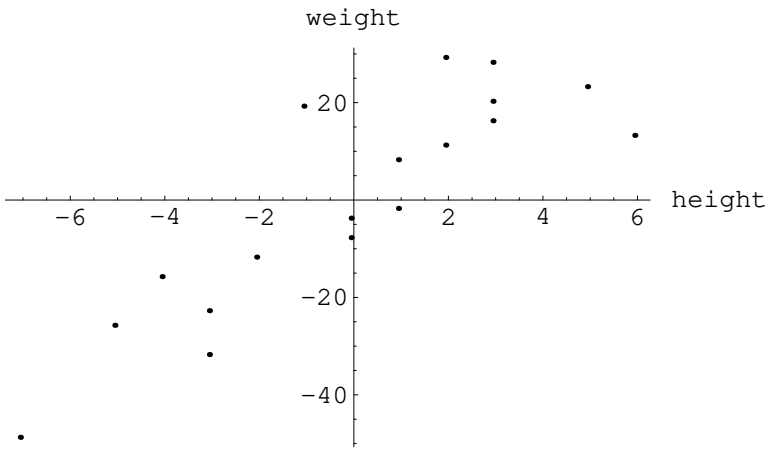
correlation matrix

```
(corhtwt = Table[acvhtwt[[i, j]] / Sqrt[acvhtwt[[i, i]] acvhtwt[[j, j]]],
    {i, Length[acvhtwt]}, {j, Length[acvhtwt]}]) // TableForm
```

```
1.           0.872905
0.872905     1.
```

■ plotting and least square fit

```
scatplot = ListPlot[nhtwt, AxesLabel → {height, weight}]
```



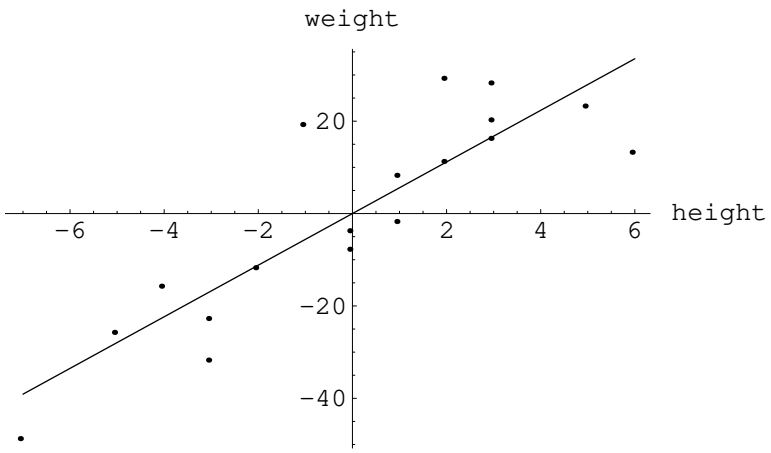

```
- Graphics -
```

least squares fit

```
lsqft = Fit[nhtwt, {1, x}, x]
```

$$1.66533 \times 10^{-15} + 5.58573\, x$$

```
Show[ Plot[lsqft, {x, -7, 6}], scatplot, AxesLabel → {height, weight}]
```



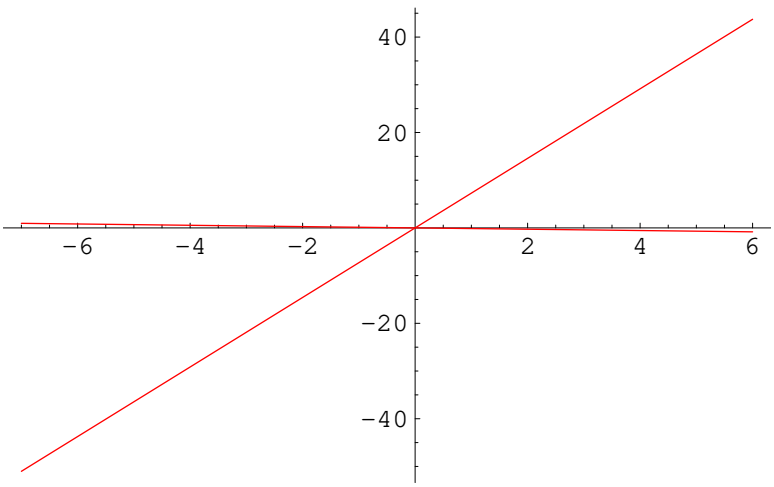

```
- Graphics -
```

Trivial PCA

```
(eightst = Eigenvectors[acvhtwt]) // TableForm
```

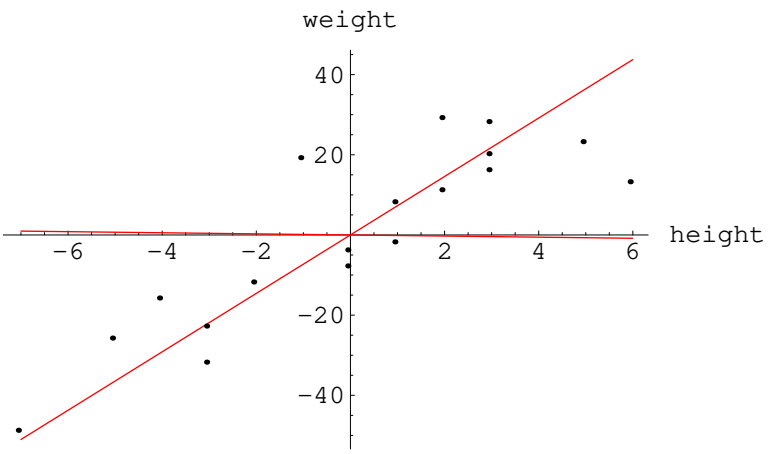
```
-0.135922      -0.99072
-0.99072       0.135922
```

```
pcaplt = Plot[
  {eightst[[1, 2]] / eightst[[1, 1]] x, eightst[[2, 2]] / eightst[[2, 1]] x},
  {x, -7, 6}, PlotStyle → {RGBColor[1, 0, 0]}]
```




```
- Graphics -
```

```
Show[pcaplt, scatplot, AxesLabel → {height, weight}]
```




```
- Graphics -
```

Why PCA?

What is it?

- **data transformation technique that preserves the information (variance) in the data but sorts the variance into uncorrelated components (axes) where the first component captures the dimension with the most variablility and subsequent components do the same with the remaining variability.**

- **typically most of the variability is cpatured in the first few components**

Why use it?

- **it may be possible that the components accountig for the smallest amount of the variability can be dropped from subsequent analysis**

- **further analysis may benefit from the fact that the axes are uncorrelated**

- **there may be interesting relationships between the original axes (variables) and the axes of the pricipal components**

Example: data in 3–space

The original data

```
(tripdata = Import["data3orig", "Table"]) // TableForm
```

```
1      2      3
2      3      4
3      1      1
4      6      7
5      5      5
6      4      2
7      8      9
8      9      8
9      7      6
```

```
tdavgs = Sum[tripdata[[i]], {i, Length[tripdata]}] / 9
```

```
{5, 5, 5}
```

```
(tdvars = Sqrt[Sum[(tripdata[[i]] - tdavgs)^2, {i, Length[tripdata]}] /
     (Length[tripdata] - 1)]) // N
```

```
{2.73861, 2.73861, 2.73861}
```

The standard scores

```
(trip = Table[(tripdata[[i]] - tdavgs) / tdvars, {i, Length[tripdata]}]) //
  TableForm // N
```

```
-1.46059      -1.09545      -0.730297
-1.09545      -0.730297     -0.365148
-0.730297     -1.46059      -1.46059
-0.365148     0.365148      0.730297
0.            0.            0.
0.365148      -0.365148     -1.09545
0.730297      1.09545       1.46059
1.09545       1.46059       1.09545
1.46059       0.730297      0.365148
```

The autocovariance matrix

```
(acvtrip =
   Sum[Dot[Transpose[{trip[[i]]}], {trip[[i]]}], {i, Length[trip]}] /
     Length[trip] // N) // TableForm
```

```
0.888889     0.740741     0.518519
0.740741     0.888889     0.814815
0.518519     0.814815     0.888889
```

The correlation matrix

```
(cortrip = Table[acvtrip[[i, j]] / Sqrt[acvtrip[[i, i]] acvtrip[[j, j]]],
    {i, Length[acvtrip]}, {j, Length[acvtrip]}]) // TableForm
```

```
1.          0.833333     0.583333
0.833333    1.           0.916667
0.583333    0.916667     1.
```

eigenvalues of the autocovariance matrix

```
(evltrip = DiagonalMatrix[Eigenvalues[cortrip]]) // TableForm
```

```
2.56364     0            0
0           0.421364     0
0           0            0.0150003
```

eignevectors of the autocovariance matrix

```
(evctrip = Transpose[Eigenvectors[cortrip]]) // TableForm
```

```
0.542129     0.761988       0.354218
0.620872     -0.0791834     -0.779903
0.566228     -0.642732      0.516024
```

```
Sqrt[evltrip] // TableForm
```

```
1.60114     0           0
0           0.649126    0
0           0           0.122476
```

component loadings

correlations between each variable and each component

```
(ldg = Dot[evctrip, Sqrt[evltrip]]) // TableForm
```

```
0.868023     0.494626      0.0433832
0.9941       -0.0514       -0.0955192
0.906608     -0.417214     0.0632005
```

proportion of the variance of each variable expladined by each compononet

```
(ldgsq = ldg ^ 2) // TableForm
```

```
0.753463     0.244655       0.0018821
0.988234     0.00264196     0.00912392
0.821938     0.174068       0.0039943
```

transform the orignal data points

```
(trntrip = Dot[trip, evctrip]) // TableForm
```

```
-1.88548      -0.556828     -0.039879
-1.25405      -0.542196     -0.00689164
-2.12979      0.497948      0.126735
0.442267      -0.776538     -0.0372718
0.            0.            0.
-0.649024     1.01123       -0.151154
1.90308       -0.469035     0.158045
2.12099       0.0149828     -0.185818
1.45201       0.820434      0.136234
```

variances of the transformed data

```
Sum[Dot[Transpose[{trntrip[[i]]}], {trntrip[[i]]}], {i, Length[trntrip]}] /
  Length[trntrip] // TableForm
```

$$\begin{array}{lll} 2.27879 & -5.42776\times10^{-16} & 4.62593\times10^{-16} \\ -5.42776\times10^{-16} & 0.374546 & -4.62593\times10^{-18} \\ 4.62593\times10^{-16} & -4.62593\times10^{-18} & 0.0133336 \end{array}$$