Pattern Recognition 2015
Unsupervised Learning

Ad Feelders

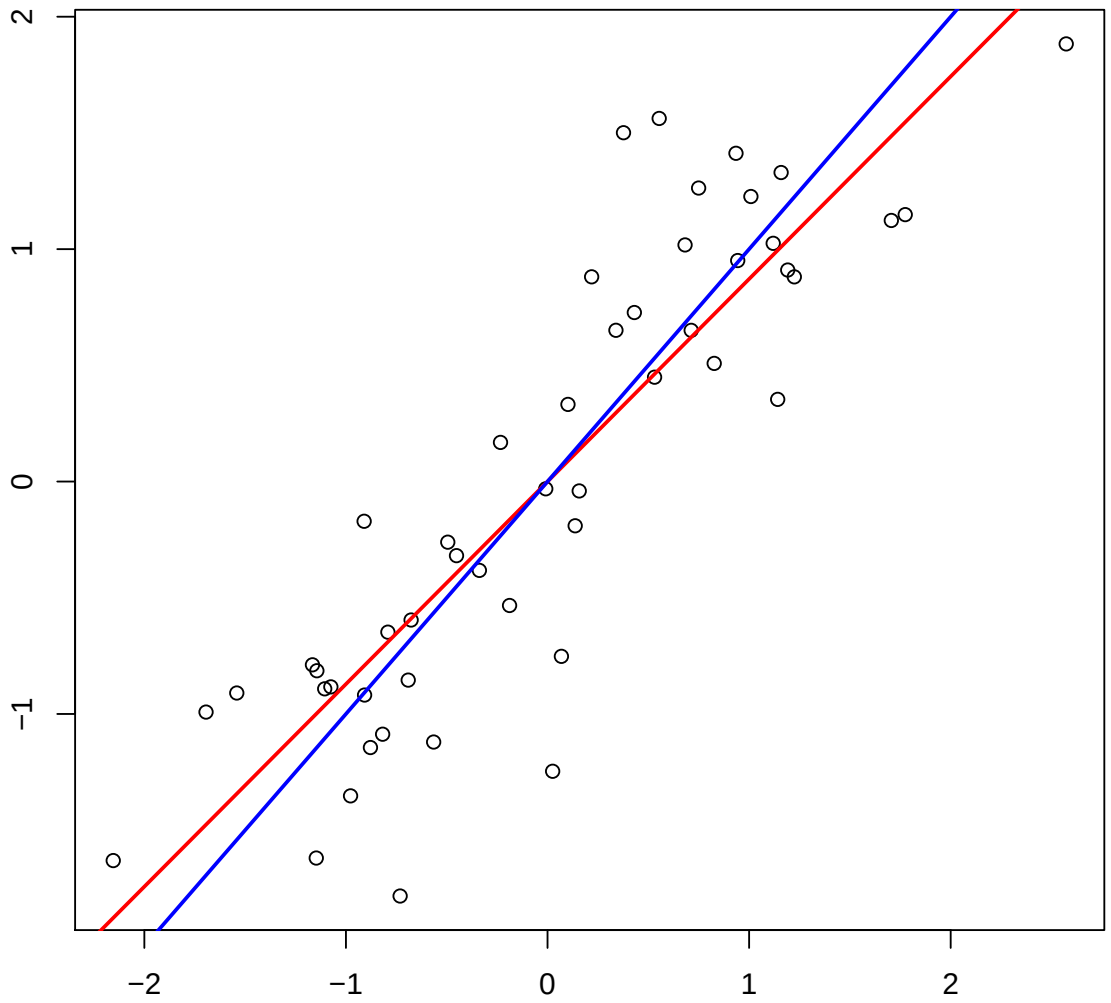
Universiteit Utrecht

Linear regression vs PCA

```
> x <- mvrnorm(50,mu=c(0,0),Sigma=matrix(c(1,.8,.8,1),nrow=2,ncol=2))
> x <- scale(x)
> plot(x[,1],x[,2])
> x.lm <- lm(x2~x1,data=data.frame(x1=x[,1],x2=x[,2]))
> abline(x.lm$coef,lwd=2,col=2)
> x.pca <- prcomp(x)
> x.pca
Standard deviations:
[1] 1.368001 0.358571

Rotation:
           PC1        PC2
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
> abline(c(0,1),lwd=2,col=4)
```

Linear regression vs PCA



Red = Linear Regression, Blue = First Principal Component

How to in R: analysis of MNIST data

```
# exclude class label and features that are always zero

> mnist.colsum <- apply(mnist.train[,-1],2,sum)
> index.colsum0 <- c(2:785)[mnist.colsum==0]

# compute principal components for MNIST training set

> digits.pca <- prcomp(mnist.train[,-c(1,index.colsum0)],scale=TRUE)

# just checking dimension of matrix with principal component scores

> dim(digits.pca$x)
[1] 42000   708
```

How to in R: analysis of MNIST data

```
# fit multinomial logit on first ten principal components
> digits.pca10.multinom <- multinom(label ~.,
  data=data.frame(cbind(digits.pca$x[,1:10],label=mnist.train[,1])),maxit=500)
# weights:  120 (99 variable)
initial  value 96708.573906
iter  10 value 30974.300003
iter  20 value 30657.503185
iter  30 value 30594.745551
iter  40 value 30544.955376
iter  50 value 30489.714174
iter  60 value 30404.840966
iter  70 value 30002.849323
iter  80 value 28599.125481
iter  90 value 28464.460457
iter 100 value 27487.669471
iter 110 value 25827.422888
final  value 25825.111402
converged
```

How to in R: analysis of MNIST data

```
> digits.pca10.pred <- predict(digits.pca10.multinom,
                       data=data.frame(digits.pca$x[,1:10]), type="class")
> table(mnist.train[,1],digits.pca10.pred)
   digits.pca10.pred
       1    2    3    4    5    6    7    8    9   10
  0 3617    1   57   45    6  248   73   10   65   10
  1    0 4408   65   38    6   75   17    5   67    3
  2   76   67 3263  206   67   14  250   77  142   15
  3   54   74  196 3351   16  279   40   71  210   60
  4   17   54   39    3 3113   36  108   66   66  570
  5  146   55   56  420  138 2609   96   18  149  108
  6   83   97  109    0   72   92 3622   16   46    0
  7   45   63   50   28   49   11    4 3716   85  350
  8   43  146  144  197   45  240   23   18 2999  208
  9   50   59   22   54  551   47    2  372  110 2921

> sum(diag(table(mnist.train[,1],digits.pca10.pred)))/42000
[1] 0.8004524

# approximately 80% correct with first ten principal components (in-sample!)
```